A SNP-based method for determining the origin of MRSA isolates

James Sciberras

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

Department of Physics

September 2015

Declarations

This dissertation is the result of my own work and includes work done in collaboration which is specifically indicated here and in the text.

Chapter 5:

- Section 5.2.4: Bayesian classification for determining an isolate's geographic origin. The method presented in this section was developed by Jukka Corander of the University of Helsinki and Richard James of the University of Bath, and implemented by James Sciberras.

Acknowledgements

I would like to thank my supervisors Richard James and Nick Priest for their continued support, encouragement and advice, both academic and personal, throughout my PhD. I would like to mention a very special thank you to Edward Feil who has also helped me immeasurably in the development of this thesis and contents therein.

I would like to thank far too many people to reasonably have space, so I will keep it brief and offer a special thank you to:

Dave Mlynski for being a sounding board for my many random ideas and discussions, as well as my comrade-in-arms throughout this whole PhD.

Sion Bayliss, Harry Thorpe, Kate Ashbrook, Ross Mounce and Tjibbe Donker for helping me understand the various processes and programs which enabled me to expand and develop this thesis;

The denizens of the Biodiversity Lab for the many stimulating, and often obscure, discussions that made my experience at the University of Bath that much sweeter;

My family and partner for their continued support, both emotionally and financially, without which none of this would have even be possible;

And finally to the University of Bath for providing a fantastic atmosphere for conducting research and providing me with the funding to pursue it.

Abstract

The advancements in Whole Genome Sequencing (WGS) have increased the amount of genomic information available for epidemiological analyses. WGS opens many avenues for investigation into the tracking of pathogens, but the rapid advancements in WGS could soon lead to a situation where traditional analytical techniques might become computationally impractical. For example, the traditional method to determine the origin of an isolate is to use phylogenetic analyses. However, phylogenetic analyses become computationally prohibitive with larger datasets and are best for retrospective epidemiology. Therefore, I investigated if there might be less computationally demanding methods of analysing the same data to obtain similar conclusions. This thesis describes a proof-of-principle method for evaluating if such alternative analysis techniques might be viable. In this thesis Methicillin resistant *Staphylococcus aureus* (MRSA) was used, and single nucleotide polymorphism (SNP) and insertion/deletion (indel) genomic variation. I move away from whole genome analysis techniques, such as phylogenetic analysis, and instead focus on individual SNPs.

I showed that genetic signals (such as SNPs and indels) can be utilised in novel ways to rapidly produce a summary of the possible geographic origin of an isolate with a minimal demand on computational power. The methods described could be added to the suite of analytical epidemiological tools and are a promising indication of the viability of developing cheap, rapid diagnostic tools to be implemented in healthcare institutions. Furthermore, the principles behind the development of the methods described in this thesis could have much wider applications than just MRSA. This implies that further work based on the principles described in this thesis on alternative pathogens could prove to be promising avenues of investigation.

List of abbreviations used.

Abbreviation	Description
ACME	Arginine catabolic mobile element
agr	Accessory gene regulator
AMPs	Antimicrobial peptides
BDOV	Bayesian Diagnostic Origin Value
BSAC	British Society of Antimicrobial Chemotherapy
CA-MRSA	Community-associated MRSA
CC	Clonal complex
ccr	Cassette chromosome recombinases
CI	Candidate Introduction
DMSS	Data Mining Surveillance System
d*N*	The number of non-synonymous nucleotide changes per non-synonymous site
DOV	Diagnostic Origin Value
d*S*	The number of synonymous nucleotide changes per synonymous site
EMRSA	Epidemic MRSA
EoE	East of England
FST	Fixation index
GWAS	Genome-wide association studies
HA-MRSA	Hospital-associated MRSA
HGT	Horizontal gene transfer
Indel	Insertion/Deletion
LA-MRSA	Livestock-associated MRSA
LSS	Location Specific SNP
MAF	Minor allele frequency
MGE	Mobile genetic element
ML	Maximum Likelihood
MLST	Multilocus sequence typing
MP	Maximum Parsimony
MRCA	Most recent common ancestor
MRSA	Methicillin-resistant *Staphylococcus aureus*
MSSA	Methicillin-susceptible *Staphylococcus aureus*
NCBI	National Centre for Biotechnology Information
NGS	Next generation sequencing
NHGRI	National Human Genome Research Institute
NJ	Neighbour Joining
OMIM	Online Mendelian inheritance in Man
PBP2a	Penicillin-binding protein 2a
PFGE	Pulse-field gel electrophoresis
PVL	Panton-Valentine Leukocidin
RC	Referral Cluster

RM systems	Restriction and modification systems
SARS	Severe acute respiratory syndrome
SCC	Staphylococcal cassette chromosomes
SEM	Standard error of the mean
SLS	Set of Linked SNPs
SnAPO	SNP-based assignment of pathogen origin
SNP	Single nucleotide polymorphism
ST	Sequence type
TAPO	Tree-based assignment of pathogen origin
UK&I	England, Scotland, Wales, Northern Ireland and Republic of Ireland
WGS	Whole genome sequencing

List of the datasets, and various subsets of the datasets, used throughout this thesis. There are two datasets which are partitioned in this thesis; a SNP dataset and an indel dataset.

Name	Isolates	SNPs / indels	Chapter	Description
Unmodified Dataset	1022	29 651	2	All isolates with all SNPs collated from the BSAC and EoE collections.
Prime Dataset	1022	29 627	2	All isolates with those SNPs left after removing the ones which might be possible homoplasies due to antibiotic resistance.
Bi-allelic Dataset	1022	5469	All	All isolates with the non-singleton, bi-allelic SNPs with no confirmed antibiotic resistance association.
Rare SNP Dataset	1022	2655	2	All isolates with the non-singleton, bi-allelic SNPs that are only expressed in two isolates.
Comparison Subset	932	5469	4	The isolates between sampled between 2001 and 2009, with the non-singleton bi-allelic SNPs.
CI Test Subset	127	5469	4	The isolates identified in Chapter 3 as possible Candidate Introduction events at the hospital geographic resolution, with the non-singleton bi-allelic SNPs.
2010 Test Subset	90	5469	4, 5, 6	The isolates sampled in 2010 with the non-singleton bi-allelic SNPs.
Indel Dataset	1009	899 indels	6	The indel dataset comprises of 899 unique indel positions across 1009 isolates.
Combined Dataset	1009	5469 SNPs, 899 indels	6	The combined SNP and indel data for those isolates which have both
Attenuated Subsets	Variable	5469	6	Iteratively removing the isolates sampled in each year, starting from 2001.
Robustness Test Subset	832	5469	6	100 isolates were removed at random from those sampled between 2001 and 2009 and used these isolates to predict the origin location of the 2010 Test Subset isolates.
2011 & 2012 Test Subset	17	5469	6	Isolates from 2011 and 2012 were used in a blind fashion to investigate if SnAPO or the Bayesian approach could determine the origin.

Table of Contents

1 Introduction

1.1 Introduction to epidemiology

Epidemiology is the study of the aetiology, transmission and identification of outbreaks of a disease. This information may then be used to inform public health policy. A disease is defined as endemic if it is maintained in a particular population (for example the distribution of *Plasmodium falciparum* malaria; Snow *et al*., 2005) or epidemic if there is a rapid spread within a short period of time (for example, the 1918 flu epidemic; Patterson & Pyle, 1991). If an epidemic is seen globally this is called a pandemic (for example, the outbreak of severe acute respiratory syndrome (SARS) in November 2002; Chan-yeung & Xu, 2003).

For many centuries, diseases were thought to have been caused by bad airs, so called "miasma". This led to such practices as blood-letting and trepanning (Plinio, 1995). Until the 19th century there was a divide in the medical community between the miasmatists who believed that diseases were propagated via bad airs, and the contagionists who believed the diseases propagated through physical contact. One of the first examples of quarantine began in the early 19th century with the segregation of smallpox and fever patients (Woodward, 1974). It was not until fairly recently that investigative epidemiology began. There were a few pivotal diseases and epidemiologists within the last couple of centuries which helped further the understanding of epidemiology and through the years the advent of microbiology, epidemiology and phylogenetics has provided insights into how infectious agents grow, transmit and evolve. However, there is still difficulty when attempting to identify the exact origin of a particular infection, which limits our ability to create treatments specific for particular outbreaks. The goal of this thesis is to provide alternative methods to the traditional techniques to attempt to determine the origin of a pathogen.

1.1.1 A brief history of modern epidemiology

John Snow is known as the father of modern epidemiology from his identification of the source of a cholera epidemic in 1854 in Broad Street, London (Johnson, 2004). However, the bacterium responsible for cholera was first reported by Gabriel Pouchet, and the

significance of its pathogenicity first realised by Filippo Pacini in 1854 (Bentivoglio & Pacini, 1995). Other important pioneers from the early days of modern epidemiology include Peter Anton Schleisner, who prevented a neonatal tetanus epidemic in 1849 (Garðarsdóttir & Guttormsson, 2009), and Ignaz Semmelweis, who used a disinfecting procedure to reduce rates of infant mortality in 1847 (Wyklicky & Skopec, 1983). However, it was not until Joseph Lister, basing his work on that of Louis Pasteur, that disinfecting became a routine practice (Bankston, 2004). The application of mathematical modelling to epidemiology was initiated by Bernoulli's work on cowpox inoculation (Bacaër, 2011). In the early 20th century mathematical models were explicitly developed for epidemiology by Janet Lane-Claypon, Anderson Gray McKendrick, Ronald Ross, and others (Heyde *et al*., 2001). The integration of mathematical models into epidemiology is one of the cornerstones of this field of research.

The latter half of the 20th century saw many instances of disease identification and acceptance. For example, the identification of the protozoan *Giardia lamblia* as the cause of giardiasis was only fully accepted in the 1970s (Ali & Hill, 2003), and the discovery in 1983 by Marshall and Warren of the bacterium *Helicobacter pylori* as the cause of certain stomach ulcers (Marshall & Warren, 1984). Even though there were reports of pathogenic enterotoxigenic *Escherichia coli* as early as 1947 (Ruchman & Dodd, 1947), there was reluctance in the scientific community to accept that *E. coli* could be pathogenic as it is a fairly common commensal of humans. It was not until the 1980s that official acceptance occurred (Riley *et al*., 1983). Another example of disease identification was the epidemiological tracking of the HIV/AIDS pandemic which started in the 1980s with the identification of a group of men who have sex with men presenting with rare opportunistic infections (Merson, 2006). Eventually, investigations led to the identification of a zoonotic origin of the virus in chimpanzees (Keele *et al*., 2006).

Modern epidemiology has seen the emergence of novel infectious agents, and the re-emergence of pre-existing pathogens (Hethcote, 2000). For example, in the 20th century there has been the identification of the causative agent for Lyme disease in 1975 (Borchers *et al*., 2014), Legionnaire's disease in 1976 (Fraser *et al*., 1977), toxic-shock syndrome in 1978 (Todd *et al*., 1978), hepatitis C & E in 1989 and 1990 respectively (Choo *et al*., 1989; Khuroo, 2011), and hantavirus pulmonary syndrome in 1993 (Khan *et al*., 1996). Furthermore, antibiotic resistant strains of a number of diseases have arisen, such as pneumonia (Breiman *et al*., 1994) and tuberculosis (Cohn *et al*., 1997). The process of identifying new infectious agents continues, and requires continually updating novel epidemiological methods. The advancement of biomedical sciences has led to the identification of certain molecular traits

that can be used as predictors of risk for a certain disease. The epidemiological study of these disease biomarkers became broadly known as “molecular epidemiology” (Maslow *et al*., 2015). Recently, genome-wide association studies (GWAS) are becoming increasingly common and are used in conjunction with epidemiology to identify genetic risk factors for many diseases and health conditions (Hirschhorn & Daly, 2005).

The recent advancement of Next Generation Sequencing (NGS) and Whole Genome Sequencing (WGS) methodologies has had a major impact on molecular epidemiology. NGS, which is also known as high-throughput sequencing, allows DNA and RNA to be sequenced quickly and cheaply. This is done by the parallelisation of the sequencing process, which allows thousands of sequences to be processed concurrently (Hall, 2007; Tucker *et al*., 2009). WGS is a process which determines the entire genome of an organism at a single time. Earlier approaches used shotgun-sequencing, whereby DNA is randomly broken up into small sequences and the overlapping regions of these small sequences are identified, and computer programs are used to reassemble the short sequences into a longer sequence (Anderson, 1981; Staden, 1979). In this way a map of the whole genome can be built, though for large genomes this could take a long time (Anderson, 1981; Staden, 1979). At present, there are a few methods being employed to accomplish cost-effective high-throughput sequencing, such as nanopore technology or pyrosequencing. The nanopore consists of a hole with an internal diameter of 1nm. When immersed in a conducting fluid, and a voltage applied, current can be observed through the conduction of ions via the pores. This current is highly sensitive to the size and shape of the nanopore. Therefore, if single nucleotides or strands of DNA pass through or near the nanopore there is a characteristic change in the magnitude of the current (Clarke *et al*., 2009). Pyrosequencing relies on the detection of pyrophosphate release on nucleotide incorporation. The intensity of the light emitted by this incorporation is directly related to the number of repeated nucleotide present in a row. This process is repeated for each of the four nucleotides, until the DNA sequence is determined (Ronaghi *et al*., 1996; Ronaghi *et al*., 1998). The dataset utilised for the development of the novel methods in this thesis was obtained through modern WGS techniques such as these.

There is a great drive to commercialise WGS, with a number of companies (e.g. Illumina, 454 Life Sciences, and Complete Genomics) competing to develop a commercially robust full genome sequencing platform. This competition, along with lucrative incentive, has driven the cost of WGS down at a rate faster than that predicted by Moore’s law (Figure 1.1). This decrease in cost has opened many avenues of investigation (e.g. personalised medicine; Lu *et al*., 2014), which may have been previously prohibitively expensive.

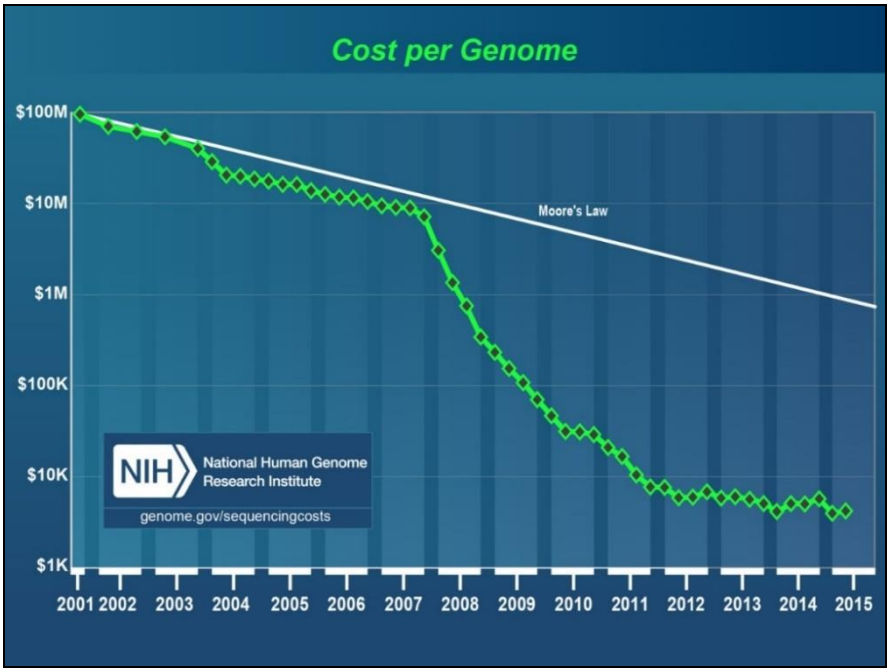


Figure 1.1. The cost to sequence a human genome has decreased drastically since 2001. This rapid decrease is faster than that predicted by Moore's law. This decrease in cost has revolutionised genomics with large quantities of data now available for analysis and interpretation at a fraction of the cost of previous years. The figure was taken with permission from the National Human Genome Research Institute (NHGRI) and is accurate as of April 2015.

Advancements in WGS have been pivotal in obtaining large quantities of data, for example, in microbiology epidemiological studies a large number of isolates may be sequenced, each of which can vary in genome size from approximately 580 kb to roughly 7 Mb (Mira *et al*., 2001). This translates into a wealth of data that requires novel techniques and methodologies to interpret in order to produce answers to the specific epidemiological question being asked. It is important to note that WGS itself does not analyse or provide interpretation of the data.

One of the prime goals of epidemiology is identifying the origin of a pathogen. This would inform as to whether a pathogen is spreading, the rate at which it is spreading, and enables the targeting of the limited resources more effectively. Therefore, in this thesis I will describe the development of novel methods which attempt to determine the possible geographic origin of a pathogen. To answer this question, I will investigate the epidemiology of methicillin-resistant *Staphylococcus aureus* (MRSA) in the United Kingdom and Republic of Ireland. The main conclusion of this work is that by adopting methodologies that permit deeper analysis of the rich diversity of information provided by whole genome sequencing one could better understand the transmission and evolution of pathogenic microbes.

1.2 *Staphylococcus aureus*

Staphylococcus aureus is a common commensal Gram-positive, facultative anaerobic, coccal bacterium of mammals. It reproduces asexually, such that it is common to refer to each separate strain as a "clone". *Staphylococcus* was first identified by Sir Alexander Ogston in 1882 (Ogston, 1882) and later differentiated into *S. aureus* and *S. albus*, now *S. epidermis*, by Friedrich Julius Rosenbach in 1884 (Licitra, 2013). Macroscopically (Figure 1.2a), the bacterium presents a golden colour due to the expression of the carotenoid pigment staphyloxanthin (Pelz *et al*., 2005). This pigment acts as an antioxidant and helps the bacterium to resist hydrogen peroxide and hydroxyl radicals, which are important in neutrophil killing (Clauditz *et al*., 2006). Microscopically (Figure 1.2b), *S. aureus* presents as clusters due to the cell division occurring in three alternative perpendicular planes. This results in irregular clumping of attached sister cells after division (Tsompanidou *et al*., 2011).

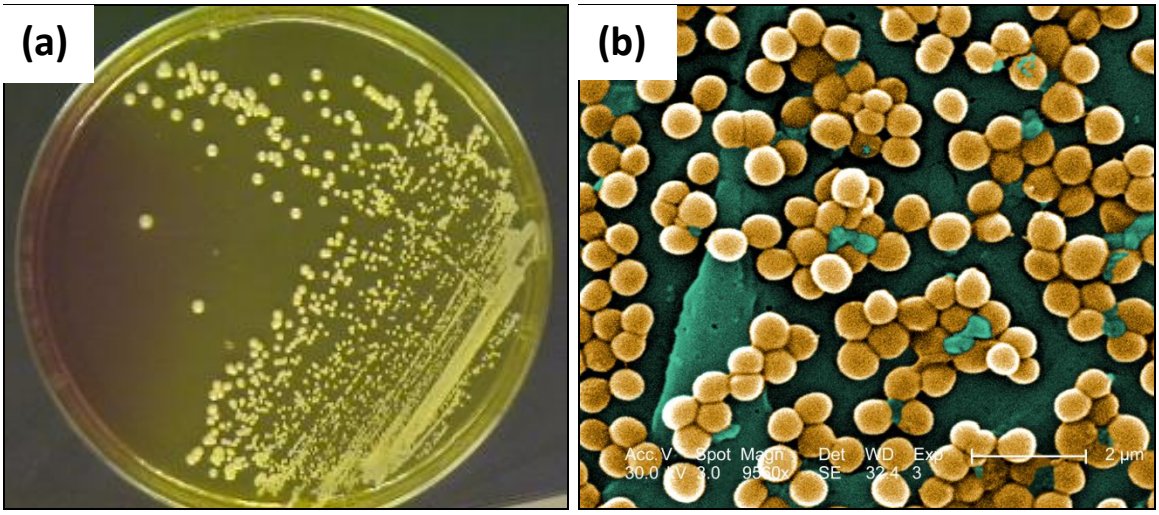


Figure 1.2. The macroscopic structure of *S. aureus* ((a); taken from www.medchrome.com) showing the characteristic golden pigmentation. The "bunch of grapes" microscopic clustering structure which gives *Staphylococcus* its name is shown in (b) (taken from www.cdc.gov).

Although primarily found in a host, *S. aureus* is robust with the ability to survive and reproduce outside a host in a variety of conditions. It can grow in harsh conditions; in temperatures between 7 to 48.5^{o}C (Schmitt *et al*., 1990), in a pH range of 4.0 to 10.0, and in environments with a sodium chloride concentration of up to 3 M (Prescott *et al*., 2005). Although it does not have flagella, it has been shown to be able to spread across soft agar using teichoic acids and surfactant molecules (Kaito & Sekimizu, 2007; Tsompanidou *et al*., 2011). This ability may help the bacterium colonise host tissue surfaces or fomites, such as medical workers' ties (Koh *et al*., 2009).

In humans the dominant niche for *S. aureus* is the non-ciliated, keratinized epithelium of the anterior nares (*vestibulum nasi*) (Kluytmans & Verbrugh, 1997; Peacock *et al*., 2001), with occasional presence in the axillia and perineum (Williams, 1963). Presence of *S. aureus* in the pharynx may be transient due to the poor adherence to nasopharyngeal ciliated epithelium (Shuter *et al*., 1996). However, the nares is likely the main reservoir for infection since colonisation of other sites may be prevented by the elimination of nasal carriage (Reagan *et al*., 1991). Due to this, the nasal carriage of *S. aureus* is a known risk factor in the development of *S. aureus* infection (Peacock *et al*., 2001; von Eiff *et al*., 2001). Nasal colonisation may be modulated by expression of secretions containing antimicrobial peptides (AMPs), such as lysozyme, lactoferrin, phospholipase A2 and defensins (Kaliner, 1991). Patients who are also carriers have a higher chance of symptomatic infection (Luzar *et al*., 1990; Weinstein, 1959) with the majority of infecting strains identical to that of the carriage strain (Nguyen *et al*., 1999; Yu *et al*., 1986).

Around 50% of individuals are nasal carriers of *S. aureus*, with 20% showing persistent carriage and the remaining 30% intermittent carriage (Kluytmans & Verbrugh, 1997; Wertheim *et al*., 2005; Williams, 1963). Current classification of nasal carriage is either "persistent" carriage or "other", and is based on bacterial load and antibody profiles. A human colonisation model shows that *S. aureus* strains have a higher survival rate in persistent carriers compared to intermittent or non-carriers (van Belkum *et al*., 2009). However, prevalence and rate of re-colonisation of *S. aureus* varies greatly and may be influenced by an individual's demographics, such as ethnicity, sex, age, or presence of a chronic illness (Cole *et al*., 2001; Lipsky *et al*., 1987; Peacock *et al*., 2003; Yu *et al*., 1986).

The exact reason why certain individuals show persistent carriage is not fully understood and is likely to be a combination of both host and bacterial factors, and the presence of other commensal organisms (Wertheim *et al*., 2005). For example, it was recently found that a serine protease, *Esp*, secreted by *S. epidermidis* led to the inhibition of *S. aureus* biofilm formation and nasal colonisation (Iwase *et al*., 2010). There is variation of *S. aureus* carriage by the host's ethnicity, suggesting a genetic predisposition of the host to colonisation (Noble, 1974). This could be through the downregulation in nasal secretions of AMPs, which may lead to increased nasal carriage (Cole *et al*., 1999). In certain individuals, haemoglobin has been associated with increased colonisation (Pynnonen *et al*., 2011). It was found that the alpha and beta chains of haemoglobin inhibit exotoxin production in *S. aureus* by downregulating the global regulator *agr* (Schlievert *et al*., 2007). This decreased virulence may allow for easier colonisation. Finally, polymorphisms in the host's immune genes, such as

glucocorticoids, C-reactive proteins, interleukin 4 and complement inhibitor proteins, may contribute to persistent nasal carriage state (Emonts *et al*., 2008; van den Akker *et al*., 2006).

S. aureus infection arises from a break in the skin or mucosa which allows bacteria to invade the bloodstream and tissues (Gordon & Lowy, 2008; Lowy, 1998). *S. aureus* is the leading cause of nosocomial infections and the list of pathophysiological complications associated is long and varied (Klein *et al*., 2007). For example, *S. aureus* is the primary cause of lower respiratory tract and surgical site infections (Richards *et al*., 1999; Richards *et al*., 2000), and one of the most common causes of bacteraemia (Naber, 2009) and pneumonia (Klein *et al*., 2007). Together with the host immune system, the combination of genetic and virulence factors influences the severity of the disease progression. However, in some cases individual factors appear to be the primary cause; for example, staphylococcal scalded skin syndrome (Amagai *et al*., 2002; Ladhani, 2003), toxic shock syndrome (McCormick *et al*., 2001), and necrotic lesions of the skin or mucosa (Lina *et al*., 1999). Therefore, *S. aureus* infection is a dangerous disease and requires rapid diagnosis and treatment to prevent fatalities.

Antibiotic resistance exacerbates the pathogenicity of the infection. Penicillin resistance was observed in *S. aureus* in the UK by the end of the 1940s (Grundmann *et al*., 2006), and observed on the other side of world in Australia in the 1950s (Rountree & Freeman, 1955). Methicillin-resistant *S. aureus* (MRSA) first appeared in an isolate in 1961 (Jevons, 1961). Surgery, prolonged hospital stay, intensive care unit (ICU) admission, and antibiotic exposure were all known risks of hospital-acquired MRSA (Chambers, 2001). Furthermore, certain populations are at higher risk; healthcare workers, those with chronic illnesses, intravenous drug users, patients with in-dwelling devices (e.g. catheters) and diabetics (Lowy, 1998).

S. aureus and MRSA pathogenicity has a great impact on healthcare resources (de Angelis *et al*., 2010; Köser *et al*., 2012), with an estimated 11.74 MRSA related hospitalisations per 1000 in the USA in 2009 alone (Klein *et al*., 2013). In the latter half of the previous decade there were reports of approximately 19,000 MRSA-associated fatalities in the US annually, equal to the number due to AIDS, TB and viral hepatitis combined (Boucher & Corey, 2008; Klevens *et al*., 2007). In England and Wales in 2012 there were 292 MRSA-related fatalities (Olatunde, 2013). However, recent control measures such as hand-washing and provision of alcohol hand-gel have been correlated with a decrease in the number of MRSA cases in the UK (Allegranzi & Pittet, 2009; Rupp *et al*., 2008). Furthermore, rates of MRSA are increasingly being used as a metric for the quality of hospital hygiene (Donker *et al*., 2010; Ke *et al*., 2012).

Grundmann *et al*. (2010) found that MRSA clones display higher levels of phylogeographic clustering than their methicillin susceptible (MSSA) counterparts. One reason for this difference could be that MRSA clones have evolved more recently than MSSA clones and so have had less time to spread. Another reason could be the greater selective pressures placed on MRSA in the harsher environments of healthcare institutions, while MSSA is under much weaker selection pressure. There is variation in the MRSA policies used in different healthcare institutions (Hails *et al*., 2003), which might enable different MRSA strains to flourish under different conditions. However, the ability of the strain to survive is highly dependent on their specific genomic components.

1.2.1 The genome of *S. aureus*

The *S. aureus* genome consists of a single circular chromosome which can range from 2.7 to 3.1 million base pairs (Holden *et al*., 2010; Lindsay & Holden, 2006) and comprises three parts; the core genome, the accessory genome, and plasmids (if present). NGS and WGS technologies have been used to determine the genomes of many different *S. aureus* strains and elucidate conserved and variable genetic regions. The core genome, which comprises approximately 75% of the *S. aureus* genome, is the region of genes which are mostly conserved across all strains (Lindsay & Holden, 2006). Genes involved in essential functions, such as metabolism and survival, comprise the majority of the genes in the core genome. However, there are also some genes involved in virulence determination, such as those coding for surface proteins, adhesins, toxins and enzymes (Lindsay & Holden, 2004).

Although mostly conserved across all strains, the core genomes are not all identical. Subtle differences in genetic composition can leave genes functional, whilst possessing different functions and phenotypes (Lindsay & Holden, 2006). Therefore, *S. aureus* can be first divided into clonal complexes and subsequently into sequence types. The method used to determine sequence types is called multilocus sequence typing (MLST) and is described in Section 1.3. A sequence type (ST) is defined by the presence of particular alleles of specific housekeeping genes, while a clonal complex (CC) is defined as a group of *S. aureus* lineages where the isolates harbours identical alleles at five or more of the loci of the housekeeping genes (see Section 1.3). Isolates from differing CCs may show variability in many genes (Lindsay & Holden, 2004). Variation in genetic composition can occur in either coding or non-coding regions of the genome through single base pair changes (single nucleotide polymorphisms, SNPs; see Section 1.2.2), insertion or deletion of one or multiple base pairs (indels), or by

recombination of large blocks of genetic sequences (Castillo-Ramírez *et al*., 2012; Lindsay & Holden, 2006).

The accessory genome of *S. aureus* contains non-essential genes that are involved in virulence, resistance and other metabolic functions (Lindsay & Holden, 2006; Shittu *et al*., 2007). The accessory genome can be highly variable between strains, due to the horizontal gene transfer (HGT) of mobile genetic elements (MGEs) from one isolate to another. This exchange is one of the mechanisms by which *S. aureus* evolves. The complete set of MGEs in a genome is called a mobilome, and can be comprised of plasmids, transposons, staphylococcal cassette chromosomes (SCCs) or lysogenic phages (Frost *et al*., 2005).

Plasmids are autonomously replicating circular segments of DNA. Previously classified according to their size and incompatibility (Novick, 1987), they are now classified by the sequence of replication (*rep*) genes (Jensen *et al*., 2010). Plasmids can encode a variety of useful characteristics, such as antibiotic resistance (e.g. pT181, tetracycline resistance in *S. aureus* strain COL; Khan & Novick, 1983), heavy metal resistance (e.g. pI258 cadmium resistance; Nucifora *et al*., 1989) and exfoliative toxin B (Amagai *et al*., 2002). There appears to be association between specific lineages of *S. aureus* and specific plasmid groups conferring virulence and antibiotic resistance (McCarthy & Lindsay, 2012). The lack of ubiquitous presence of any one plasmid group in all *S. aureus* lineages indicates that there may be some impediment to the evolution of a hyper-resistant and hyper-virulent *S. aureus* strain.

SCCs are DNA fragments of 21-53kb in length. They may contain genes that affect antibiotic resistance or virulence factors. There are two main classifications of SCC; those with the *mecA* gene (SCC*mec*) which confers resistance to β-lactam antibiotics by coding for an alternative penicillin binding protein (PBP2a) which has extremely low reactivity with β-lactam antibiotics (Kim *et al*., 2012; Ubukata *et al*., 1989), and those without. The transcription and translation of the resistance protein PBP2a influences the minimum concentration of β-lactam antibiotics required to inhibit growth (Hartman & Tomasz, 1984). Additional SCC*mec* elements are *mecI* and *mecR*, which code for the repressor and signal membrane transducer respectively (Noto *et al*., 2008; Ubukata *et al*., 1989). Apart from the *mec* gene, all SCC*mec* element genes also contain cassette chromosome recombinases (*ccrA*, *ccrB*, *ccrC;* Noto *et al*., 2008). The combination of *mec* gene with *ccr* determines the classification of the SCC*mec* element, with eight identified thus far in *S. aureus* (Ito *et al*., 2013).

Transposons are usually small and encode antibiotic resistance (e.g. Tn*554* encodes for erythromycin resistance; Phillips & Novick, 1979). They contain a transposase gene (Rowland &

Dyke, 1989) which allows them to excise, replicate and integrate into a chromosome or other MGEs, such as SCCs or plasmids (Baba *et al*., 2002).

There are three processes by which HGT occurs to transfer MGEs; transformation, conjugation and transduction (Thomas & Nielsen, 2005). Transformation is the uptake of exogenous DNA directly from the surroundings. It was only recently that a successful transformation in *S. aureus* was observed (Morikawa *et al*., 2012). Conjugation is the transfer of genetic material during direct cell-to-cell contact via pili or pores. Due to the lack of pili in *S. aureus* it is believed that conjugation occurs through the pores (Lindsay, 2014). However this is not a common HGT process in *S. aureus* (Lindsay & Holden, 2006) since transfer (*tra*) genes are required to be present in the plasmid to be transferred (Guglielmini *et al*., 2013) and only a very few *S. aureus* isolates express these genes (McCarthy & Lindsay, 2012). Transduction is the main mechanism by which HGT occurs in *S. aureus* and requires a bacteriophage to act as a vector. A bacterium may be infected with a bacteriophage and either incorporate their genetic material into the bacterial chromosome or deliver foreign DNA (in the form of bacterial chromosomes or plasmids). This process is known as generalised transduction (Lindsay & Holden, 2006). Once incorporated the phage can either replicate and eventually lyse the bacterial cell releasing new phages (lytic phage), or remain integrated within the host genome (lysogenic phage). Many strains of *S. aureus* contain between one and four lysogenic phage types (Lindsay & Holden, 2006). Bacteriophages can carry virulence genes which code for factors such as Panton-Valentine Leukocidin (PVL; Kaneko *et al*., 1998), enterotoxin A (Betley & Mekalanos, 1985), and exfoliative toxin A (Yamaguchi *et al*., 2000).

1.2.2 Genetic variation in *S. aureus*

Apart from the assimilation of MGEs into the genome, genetic variation may arise through point mutations, indels, or recombination. Indels which are not multiples of three nucleotides cause a frameshift effect to occur in coding regions which could result in the production of a different protein chain, and would usually be subjected to purifying selection in coding regions (Chen *et al*., 2009). Indels have been used to infer phylogenetic relationships (Pereira *et al*., 2010) and as genetic markers (Väli *et al*., 2008). However, it may be difficult to infer the direction of evolution (i.e. has organism A lost a nucleotide, or has organism B gained one) through only using indels. In bacterial genomes, there is a significant deletion bias, with many more deletions than insertions identified (Mira *et al*., 2001). Homologous recombination also contributes substantially to genomic variation due to the exchange of up to several kilobases of genetic material (Castillo-Ramírez *et al*., 2012; Lindsay & Holden, 2006); for

example in the variable region of *agrB* and *agrC* in the accessory gene regulator operon (Dufour *et al*., 2002). Furthermore, there is some evidence to suggest that recombination, combined with selection, may favour the evolution of novel clonal complexes of *S. aureus* (Basic-Hammer *et al*., 2010).

However, the most abundant type of variation is point mutation; SNPs. This is the variation of a single nucleotide (**A**denine, **C**ytosine, **G**uanine or **T**hymine) in a genetic sequence. SNPs are either synonymous or non-synonymous (or neither in intergenic regions), and may result in subtle changes in the genome. With their accumulation they provide the majority of diversity between genomes (Gouy & Gautier, 1982). Due to the degeneracy in the amino acid code the majority of SNPs are synonymous ("silent") mutations and do not result in a change in the functionality of the gene expression. However, non-synonymous mutations lead to a change in the amino acid and hence a potential altered gene or protein expression. The ratio of the number of non-synonymous nucleotide changes per non-synonymous site (d*N*) and the number of synonymous changes per synonymous site (d*S*) is often used to determine the rate of evolution in, or between, organisms (Hurst, 2002). However, work by Rocha *et al*. (2006) indicates that the dN/dS ratio might not be constant for an organism, and rather one should use the trajectories of d*N*/d*S* over time. *S. aureus* has a higher non-synonymous change bias within clones (d*N*/d*S* ~ 0.7) than between clones (d*N*/d*S* ~ 0.1) and this higher level of synonymous change between clones could be attributed to the acquisition of MGEs through HGT (Castillo-Ramírez *et al*., 2011). The four nucleotides are divided into pyrimidines (C and T) and purines (A and G). Therefore, SNPs may be either transversion (purine to pyrimidine, or vice-versa) or transition (purine to purine, or pyrimidine to pyrimidine) mutations. Although there are four ways for a transversion mutation to occur, compared to two ways for a transition, approximately 2/3rd of SNPs are transition mutations (Collins & Jukes, 1994). This is since transitions do not require alteration of the molecular ring structure (e.g. a single ring purine to a single ring purine), whereas transversions are changing the number of molecular rings (e.g. a single ring purine to a double ring pyrimidine, or vice-versa).

The majority of SNPs are bi-allelic, only showing two possible nucleotides, and can be easily assayed (Sachidanandam *et al*., 2001). The frequency of the least common allele is defined as the minor allele frequency (MAF), and is usually specific to a particular population of the organism. The distribution of SNPs in a genome is often not homogenous; for example, in humans more SNPs occur in non-coding regions than those coding for specific genes (Varela & Amos, 2010). SNP density of a particular genetic region in humans can be predicted by the presence of microsatellites and the GC content (Varela & Amos, 2010). The fixing of specific

alleles in a population, mutation rate and recombination may all affect SNP density in the genome (Barreiro *et al*, 2008; Nachman, 2001). Linkage disequilibrium (i.e. the non-random association of alleles at different genetic loci) might be causing the co-occurrence of particular SNPs (Slatkin, 2008). A tag SNP may be used to represent a group of SNPs in a region of high linkage disequilibrium (Chen *et al*., 2014).

In coding regions non-synonymous mutations can either create a missense mutation, where the amino acid code is changed (e.g. Asn→Asp at position 637 in the *ileS* gene in MRSA strains SA-7S and SA-7R; Fujimura *et al*., 2003), or a nonsense mutation, where a premature stop codon is expressed (e.g. the avirulence of strain 8325-4 is attributable to a nonsense mutation in the *agrA* gene; Adhikari *et al*., 2007). Therefore, non-synonymous mutations may influence the expression of a gene and this altered expression may lead to antibiotic resistance; for example, the C→T mutation in strain ST239 at position 7255 in the DNA gyrase subunit A gene is involved in quinolone resistance (Harris *et al*., 2010).

SNPs have a wide range of applications in epidemiology; for example, the identification of SNPs in certain diseases as next generation markers for pharmacogenomic targets of drug therapy (Lai, 2001). In microbiology, SNPs are important genetic markers which gained popularity at the turn of the millennium (Vignal *et al*, 2002). They can help identify relationships between strains of a pathogen, either through direct assay or the creation of a SNP-based phylogenetic tree. Unless specifically selected against or lost through genetic drift, SNPs accumulate in a genome over time and are often stably inherited (Thomas *et al*., 2011). Therefore, knowledge of the mutation rate will help infer relationships between isolates. For example, the mutation rate of *S. aureus* strain ST239 has been estimated as 3.3×10^{-6} per site per year (Harris *et al*., 2010). Furthermore, the probability of two independent base changes at a single position is very low (Vignal *et al*., 2002), decreasing the probability of a homoplasy occurring. Therefore, SNPs can be used as a stable signal for the propagation of a particular strain. This use is extended to population genetics for estimating genetic variation, identification of relatedness or parentage, measuring population structure and changes in population size over time (Morin *et al*., 2004).

Historically SNPs were found through laboratory methods (e.g. gel electrophoresis, or restriction fragment length polymorphism), but modern day techniques use sequencing and identification *in silico* (Morin *et al*., 2004). The advancements in WGS allow for rapid sequencing of entire genomes, which can be processed *in silico* to determine the harboured SNPs. Each isolate sequenced is compared against a relevant reference genome, and the

nucleotides which deviate from the reference genome are identified as SNPs. There are a number of growing online databases to which the SNP could be added. Many of them host information of multiple species, though there is often a focus on human genomes. For example, dbSNP is hosted by the National Centre for Biotechnology Information (NCBI) and contains information on 55 different organisms (Morin *et al*., 2004); Kaviar collates SNPs from multiple data sources of human genomes to inform on personalised medicine (Glusman *et al*., 2011); SNPedia supports personal genome annotation and analysis of human genomes; OMIM database collates the known SNP-associated diseases in humans; and the Human Gene Mutation Database provides the gene mutations associated with inherited disease. The development of collated online databases is useful since SNP data, which in most cases is bi-allelic and hence a simple Presence/Absence question, can be easily compared between different laboratories, unlike the inconsistency in determining allele size in microsatellite analysis (Delmotte *et al*., 2001; Vignal *et al*., 2002). Furthermore, SNPs can provide equivalent statistical power to microsatellites, yet cover a wider range of the genome (Morin *et al*., 2004). The information on the position and effects of SNPs has considerable potential to further our epidemiological understanding of certain diseases. However, the large quantities of information available in these databases and through WGS can be difficult to interpret and therefore novel methods may be required.

1.3 Methicillin-resistant *Staphylococcus aureus*

Methicillin-resistant *S. aureus* (MRSA) was first reported in 1961, and likely began with the introduction of a SCC*mec* I element into a methicillin susceptible *S. aureus* (MSSA) isolate belonging to ST250 (Enright *et al*., 2002). The exact origin of this MGE is unknown but the evolutionary precursor of the *mecA* gene found in MRSA strains could be from *S. sciuri* (Wu *et al*., 1996; Wu *et al*., 2001). There appears to be at least 20 separate acquisitions of SCC*mec* into *S. aureus* (Robinson & Enright, 2003), indicated by strains from the same ST containing different SCC*mec* types.

Protocols have been established to deal with the presence of MRSA; such as provision of alcoholic hand-gel (Rupp *et al*., 2008), increased hand washing (Allegranzi & Pittet, 2009), increased sanitation (Dancer, 2009), and the use of the final front-line drug, vancomycin (Bassetti *et al*., 2009). These methodologies are correlated with a decline in the rates of invasive MRSA infection in the UK (Office for National Statistics, 2011). However, a main drawback is that these approaches are generalist, treating every case of MRSA in the same way. Difficulties arise when the various strains of MRSA encountered respond differently. Within MRSA isolates there is some diversity, with more than 100 strains identified by Monecke *et al*. (2011). Knowledge of the genetic characteristics and relatedness of the MRSA isolates in a particular sub-population is required for targeted, effective action. There are a number of methods developed to identify the particular genotype of an isolate.

Multilocus sequence typing (MLST) is used to determine the sequence type (ST) and clonal complex (CC) for a given isolate. MLST works by identifying which alleles are present in the *S. aureus* genome of seven different housekeeping genes (Maiden *et al*., 1998). The combination of specific alleles encodes for a specific sequence type (ST; Enright *et al*., 2000), with five or more identical alleles indicating the isolate belongs to a specific CC. Since there is low mutation rate in these housekeeping genes, MLST can be used for long evolutionary time which encompasses many generations. This is especially important in MRSA, since generational times can be measured in minutes (Laurent *et al*., 2001). Pulse-field gel electrophoresis (PFGE) separates chromosomal DNA fragments, digested by the enzyme *Sma*I, using agarose gel electrophoresis with an alternative voltage gradient. The resulting band patterns are processed using specialised software and related strains are grouped together (McDougal *et al*., 2003). One other sequence based analysis determines the variation in the polymorphic X-region of the protein A gene (*spa*; Frénay *et al*., 1996). This *spa*-typing allows for discrimination between strains based on the number of tandem repeats in the sequence. With only one locus

sequenced, this method is cheaper and faster than MLST, however there is a corresponding reduction in discriminatory power if the same *spa* gene occurs in multiple clonal lineages through recombination (Deurenberg & Stobberingh, 2008).

There are approximately 11 major *S. aureus* clonal complexes (CC1, CC5, CC8, CC12, CC15, CC22, CC25, CC30, CC45, and CC51) with many more smaller ones, and most MRSA isolates in healthcare locations belonging to CC5, CC8, CC22, CC30 and CC45 (Lindsay & Holden, 2006). In these five CCs, containing both MRSA and MSSA, there are at least 11 major STs (Enright *et al*., 2002); for example, ST247 is globally widespread but is commonly known as the Iberian clone (Sanches *et al*., 1995), and the highly transmissible and multi-drug resistant clone ST239 is known as the Brazilian clone (Johnson *et al*., 2001). These clones can be distinguished from each other through PFGE. There are a number of other major epidemic (E) MRSA sequence types that are found across the world; ST5, ST22, ST30 and ST45. Furthermore, although containing MSSA clones ST8 also contains EMRSA clones that have acquired SCC*mec* types II and IV (Enright *et al*., 2002). However, although some strains appear to dominate certain geographic locations, there is still great diversity seen at any one location. For example, Feil *et al*. (2003) found 75 unique sequence types from 334 isolates recovered from Oxfordshire, UK (Figure 1.3).

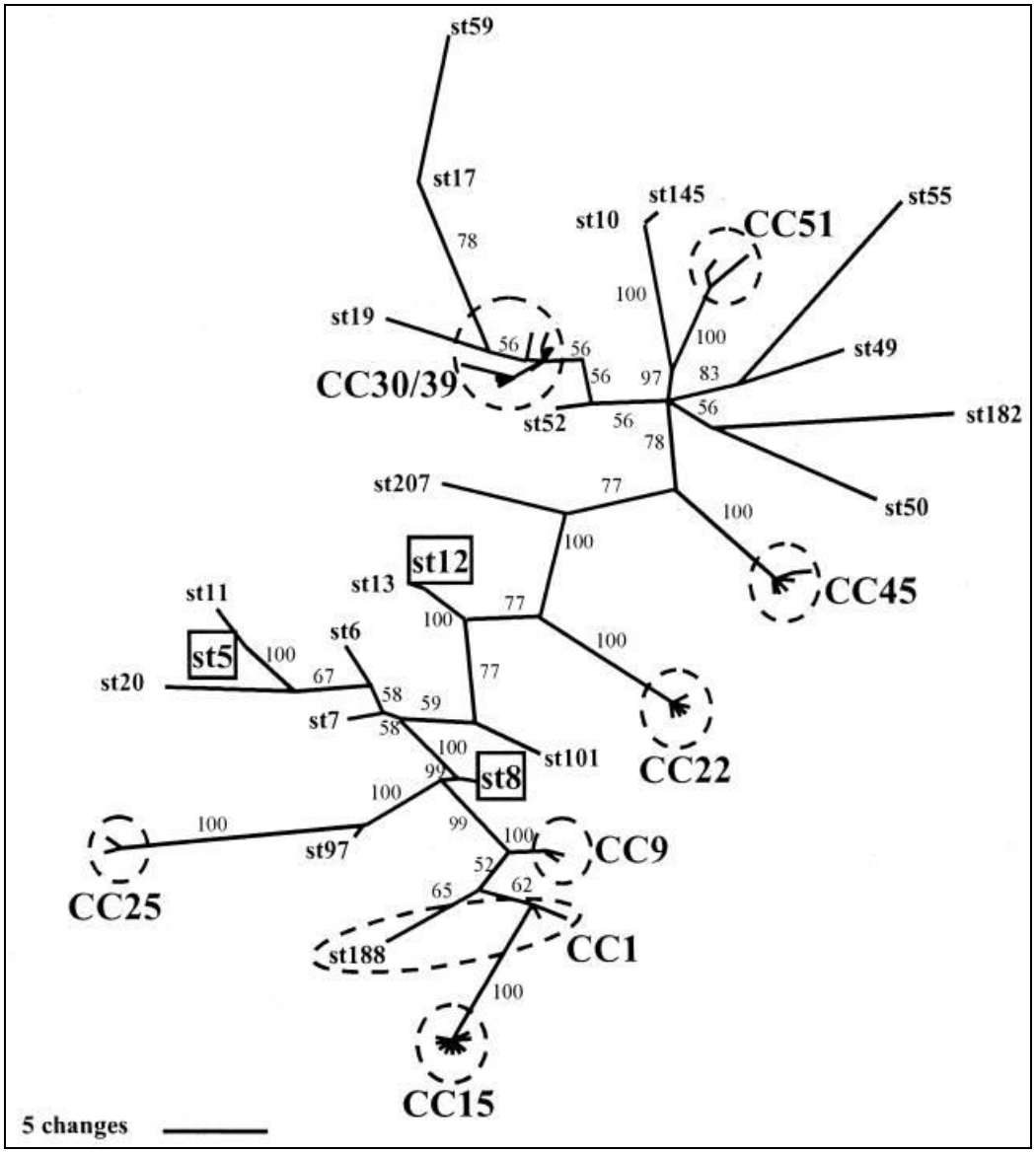


Figure 1.3. This unrooted Bayesian phylogenetic tree is taken from Feil *et al*. (2003) and shows the diversity of the 75 different STs and CCs found in Oxfordshire, UK. All the major clonal complexes are denoted by the dashed rings, while the boxes represent singletons or minor groups that comprise of more than five isolates. The individual STs within each major clonal complex are not labelled, barring ST188. The isolates within a CC appear to cluster together very tightly, with comparatively large evolutionary distance between CCs.

Successful HGT of MGEs is attributed to be the reason for the evolutionary success and dominance of certain MRSA clones (Lindsay *et al*., 2012). There is strong evidence that MGEs contribute to the emergence of highly virulent and multi-drug resistant clones; for example, a clone of the ST772 lineage shows resistance to six different antibiotics, mostly acquired through HGT of MGEs (Steinig *et al*., 2015). Furthermore, recombination plays an important role in the diversification of MGEs. Large chromosomal replacements via homologous recombination are likely to influence the long-term evolution of MRSA (Feil *et al*., 2003; Grundmann *et al*., 2006). For example, ST239 is a descendent of ST8 which acquired a large

(635kb) fragment from ST30 (Holden *et al*., 2010) and exhibits a mosaic genome profile. Furthermore, the ST239 clone also showed significant recombination variation, associated with phylogeography (Castillo-Ramírez *et al*., 2012).

There is some evidence that MGEs are transferred with higher frequency in some lineages (Lindsay, 2014). There are preventative measures to impede the development of a hyper-resistant clone, with *S. aureus* expressing restriction and modification (RM) systems which destroy certain 'foreign' DNA based on the sequence and modification patterns (Lindsay, 2014). These RM systems usually prevent transfer of virulence or resistance genes between different CCs but in certain situations the lack of a particular target site allows a genetic transfer between lineages, which may be instrumental in driving the evolution of MRSA clones (Roberts *et al*., 2013).

Unfortunately, the methods mentioned in this section only allow the categorisation of isolates to a particular lineage. Isolates within a CC would appear to be very similar and therefore cluster on a phylogenetic tree, while isolates from different CCs would show a distinct evolutionary distance. Furthermore, MSSA often shows very high genetic diversity and would also show significant evolutionary distance (Vandendriessche *et al*., 2013). It is difficult to use these methods to identify exact relatedness between any two isolates. Therefore higher genetic resolution is required. The publication of the first MRSA genome in 2001 (Kuroda *et al*., 2001), the development of Next Generation Sequencing (NGS) techniques, and advancements in Whole Genome Sequencing (WGS) has revolutionised sequencing throughput efficiency as many sequences can be processed in parallel. Furthermore, the cost for rapid WGS of one MRSA isolate is currently less than £50 (Priest *et al*., 2012). WGS and NGS techniques can be used to generate SNP and indel maps of a particular isolate when compared to a reference genome (Harris *et al*., 2010). Multiple strains can be sequenced efficiently and all deviations from the reference sequence identified. The uses of WGS and NGS with respect to MRSA are developed further in Section 1.4.2.

1.3.1 Epidemiology of MRSA

Since the first appearance of methicillin resistance in the 1960s, MRSA strains are now endemic, and epidemic, in healthcare facilities and communities across the globe (Ayliffe, 1997; Boucher & Corey, 2008; Diekema *et al*., 2004; Enright, 2003) due to rapid spread and colonisation (Roman *et al*., 1997). The population structure of MRSA is mainly clonal and it appears that point mutations, and not recombination, were predominant in the initial stages of clonal diversification (Enright *et al*., 2002; Feil *et al*., 2003). The population structure in

MRSA provides a good balance between genetic variation due in the SNPs and genetic conservation due to limited recombination, unlike other nosocomial pathogens which exhibit different genetic and population structure; for example, the many chromosomal rearrangements in the yeast *Candida albicans* (Rustchenko-bulgac *et al*., 1990) which may be found in intravascular catheters and can cause bloodstream infections (Akbari & Kjellerup, 2015).

Prevalence of MRSA is not globally uniform. The US, south-east Asia, southern Europe, parts of South America and Australia all show higher incidence rates (Grundmann *et al*., 2006), while the Netherlands and Scandinavian countries have the lowest MRSA prevalence rates. This is likely due to their highly effective search-and-destroy policies, which involves screening, quarantine, and eradication (Simoens *et al*, 2009). Although originally thought to be restricted to hospitals and other healthcare institutions there are now three epidemiological classifications of MRSA; healthcare acquired (HA), community acquired (CA), and livestock acquired (LA) MRSA. There are a number of LA-MRSA lineages; for example the CC398 lineage found in pigs, cattle and poultry (Köck *et al*., 2013), and the CC97 lineage, primarily found in cattle (Spoor *et al*., 2013).

One primarily HA clone currently disseminating in the UK and abroad is EMRSA-15, from the ST22 lineage (Johnson *et al*., 2001). This clone showed rapid and efficient transmission within the hospital environment and is one of the most common MRSA clones in Europe (Grundmann *et al*., 2010), with transmissions from the UK to many different countries, such as Portugal, Germany, Spain and many more (Holden *et al*., 2013). To understand the genetic changes that contributed to the success of this clone, Holden *et al*., (2013) sequenced 193 ST22 genomes from 15 countries and identified two non-synonymous SNP mutations associated with fluoroquinolone resistance as the key factors for the success of this particular clone. This resistance conferred a competitive advantage in environments where this antibiotic was frequently used. This study also identified, using Bayesian reconstruction, the origin of this clone to be in the middle of the UK in the mid-1980s. Furthermore, the ST22 isolates used in the Holden *et al* (2013) study shows a comet shaped Maximum Likelihood phylogeny, indicating the high level of clonal similarity in the majority of the isolates. Phylogenetic analysis in this study also indicates the single acquisition of SCC*mec* IVh just prior to the emergence of ST22. The ST22 lineage, especially the EMRSA-15 clone, is one of the main strains disseminating in UK healthcare institutions. This is one of the reasons why this clone is the one chosen to comprise the dataset used for analysis in this thesis.

CC30, the lineage to which phage type 80/81 belonged, also harbours one of the more successful HA epidemic clones, EMRSA-16 (Robinson *et al*., 2005). Isolates from both strains show that they diverged from a recent common ancestor (DeLeo *et al*., 2011). However, EMRSA-16 is less virulent than phage type 80/81 and is restricted to hospitals. This decrease in virulence could be due to the acquisition of SCC*mec* type II which, while still conferring β-lactam antibiotic resistance, can interfere with *agr* signalling and a corresponding decrease in cytolytic toxin expression (Rudkin *et al*., 2012). Another example of specific genetic characteristics conferring advantages is the successful spread of ST239 in China following the recent acquisition of the s*asX* gene, a novel cell wall-anchored protein gene which may be related to the infection invasiveness (Li *et al*., 2012). Additionally, an increase in the presence of the ACME element in ST239 strains has been discovered in Singaporean hospitals between 2006 and 2009 (Hon *et al*., 2013).

Although this thesis primarily focuses on HA-MRSA isolates, there is currently growing concern over CA-MRSA, which appears to be acquired in individuals with no prior contact with healthcare institutions (Chambers & Deleo, 2009) or any other known risk factors. CA-MRSA emerged in the late 1990s and differs in both genotype and phenotype from HA-MRSA (Todd *et al*., 2005). CA-MRSA exhibits more clonal diversity than HA-MRSA (Feng *et al*., 2008; Francois *et al*., 2008). One of the possible origins of CA-MRSA is Panton-Valentine Leukocidin (PVL)-positive MSSA strains present in Japan which acquired SCC*mec* IV (Taneike *et al*., 2006). Within a few years CA-MRSA has spread globally, with different lineages of virulent CA-MRSA strains conforming to specific geographical predominance (Mediavilla *et al*., 2012). A ST1 clone is prevalent in Asia, Europe and the US, a ST30 clone is disseminating in Australia, Europe and South America, and a ST80 clone is in Asian Europe and the Middle East (Deurenberg & Stobberingh, 2008). The USA300 clone from ST8 is found mainly in North America (Patel *et al*., 2013) with some dissemination in Europe (Witte *et al*., 2007) and South America (Reyes *et al*., 2009), and is considered to be one of the most severe outbreaks globally (Otto, 2010). USA300 appears to have emerged in the early 1990s (Uhlemann *et al*., 2014) and expresses the arginine catabolic mobile element (ACME), which is posited to be important in the overall fitness and transmissibility of the strain (Diep *et al*., 2008). There is some evidence that ACME might be responsible for attenuated virulence in certain isolates (Diep *et al*., 2008). Another reason for the dominance of USA300 in North America could be its resistance to polyamines, to which most *S. aureus* strains are hyper-sensitive (Joshi *et al*., 2011).

Unlike HA-MRSA, CA-MRSA strains appear to cause disease, and a higher rate of mortality, in younger patients with no previously defined health risk factors (Fridkin *et al*.,

2005; Hunt *et al*., 1999). CA-MRSA is highly associated with an increase in skin and soft tissue infections (SSTIs; Otto, 2010), and in rare cases can cause more severe, often fatal diseases such as necrotising pneumonia, Waterhouse-Friderichsen syndrome and necrotizing fasciitis (Klein *et al*, 2009; Miller *et al*., 2005). The increased virulence in CA-MRSA may be attributed to two possible MGEs, SCC*mec* type IV and prophage ΦSA2pvl (Baba *et al*., 2002). The SCC*mec* elements associated with HA-MRSA (SCC*mec* I, II and III) appear to have a greater fitness burden than the shorter SCC*mec* elements associated with CA-MRSA (SCC*mec* IV, V and VII; Lee *et al*., 2007). Due to the nature of their respective selective pressures, HA-MRSA is generally multi-drug resistant, while CA-MRSA isolates are usually non-beta-lactam antibiotic sensitive (Otto, 2013). Importantly, HA-MRSA requires much higher antibiotic concentrations for efficient eradication (Otto, 2010). Furthermore, non-genetic factors such as socioeconomic standards, rate of incarceration, and availability of adequate healthcare may all influence the transmission and success of CA-MRSA (Witte, 2009).

HA-MRSA historically outcompeted CA-MRSA in healthcare institutions due to expression of higher antibiotic resistance. However, currently both HA- and CA-MRSA circulate in the community, with some CA-MRSA clones invading hospitals and healthcare facilities (Witte, 2009). This leads to a more complex epidemiology, though certain molecular traits can be used to distinguish between the two groups since some traits are associated mainly with CA-MRSA; for example SCC*mec* Type IV and V, or genes encoding PVL (Deurenberg *et al*., 2007). This increase in the mixing of the MRSA clones and sub-populations warrants the development of methods which can determine the origin of an isolate, in order to facilitate treatment and inform preventative measures. The methods developed in this thesis would be applicable to CA-MRSA as well as HA-MRSA.

It has been identified that humans are the vector by which different strains of MRSA move around the country and the world (Donker *et al*., 2014; Donker *et al*., 2010; Donker *et al*., 2012; Ke *et al*., 2012; Lutz *et al*., 2014; Mendiola *et al*., 2015). There are a number of methodologies to determine relatedness of MRSA isolates and hence the spread of particular strains from patient to patient, and institution to institution (David & Daum, 2014), as described in Section 1.3.2. However, the first step is obtaining the data.

1.3.2 Epidemiological surveillance methodologies and MRSA

There are currently many different ways in which samples and information of a particular MRSA strain and its pathogenic effects can be obtained. Both observational and experimental data are used to help identify, categorise and combat a pathogenic threat.

Epidemiologists may use qualitative studies of a single or group of patients, or select specific subjects based on their disease status. Cohort studies, such as GWAS, are very important and informative in understanding how particular conditions affect a group of individuals. Outbreak studies are the investigation of a sudden increase of a disease at a particular location or time and may be identified passively or actively. The number of cases which classifies an outbreak is relevant to the disease's rarity and virulence. Outbreaks may be attributable to a common source, or propagated from individual to individual. Behavioural risks (e.g. working in a healthcare facility) or zoonotic transmissibility (e.g. LA-MRSA) may also influence the spread. Epidemiology studies are often used to inform governments and pan-governmental bodies of the risks posed by the particular MRSA strain. In essence, they aid health management in assessing the needs of a population and to efficiently implement the appropriate interventions required to limit transmission and impact.

There are two levels with which the surveillance is conducted. Passive surveillance gathers data from all reporting healthcare workers as routine, but does not specifically require that the data be reported. Thus, passive surveillance is the cheapest and most common form of surveillance. However, passive surveillance usually increases the delay before proper identification of an outbreak can occur. Active surveillance requires more resources and trained practitioners, using multiple sources of data, to detect issues earlier on (Peterson & Brossette, 2002). More active surveillance is required to combat the modern epidemiological threat. However, this usually requires routine screening, targeted drug use and trained practitioners which all cost considerable resources. Although expensive, active surveillance could prevent, or at least slow down, the dissemination of a MRSA strain by identifying the outbreak faster and therefore minimising the number of individuals at risk (Peterson & Brossette, 2002). Furthermore, strong cooperation is required between healthcare institutions and clinical epidemiologists as the pathogens of concern affect both in- and out-patients, healthcare workers and the general community (Peterson & Brossette, 2002).

The ideal surveillance system would contain analysis tools that automatically identify novel outbreaks, unusual patterns of spread, and determine the origin of the pathogen (Dean, 1994). This may involve pattern identification and data mining to survey hospital and outpatient data (Peterson & Brossette, 2002). There is considerable drive to automate the surveillance, as this appears to provide the best way to rapidly assess and detect infectious diseases in our healthcare system (Peterson & Brossette, 2002). The Data Mining Surveillance System (DMSS; Brossette *et al.*, 2000) represents the first generation of semi-automated surveillance systems for nosocomial infections. Although the DMSS appears to function well at

condensing and summarising the relevant information, this system still requires a manual interpretation by a trained investigator. However, the authors do note that this is just the first step and more work is required to develop these systems. Furthermore, an online central database could be used to monitor data from various institutions and allow rapid communication of emerging infectious diseases and be able to rapidly identify outbreaks (Bogich *et al*., 2013). Rapid identification is important, since the critical response time (i.e. the time period before the outbreak becomes an epidemic) for a given strain of MRSA may be very short (Rivas *et al*., 2003). Furthermore, reduction in identification time will reduce the time that colonised patients may have to disseminate the pathogen (Tacconelli, 2009). Ciccolini *et al*. (2014) showed that it would be possible to build an efficient early detection system for nosocomial pathogens using sentinel hospitals and the knowledge that patient referrals are the likely vector of transmission between locations.

Although some surveillance methods have been developed (Mellmann *et al*., 2006; Peterson & Brossette, 2002), the current trend for ever cheaper WGS (Harris *et al*., 2013) opens up many avenues of approach to analyse an isolate's genetic content. The increasing viability of WGS and NGS would permit their use in routine surveillance. There is a clear drive at the moment for developing greater global surveillance systems of MRSA in order to identify new introductions and outbreaks (Harris *et al*., 2010). Molecular epidemiology may generate large quantities of data, which need to be correctly analysed to identify patterns in the progression of a disease. If it was known where the isolate had originated from, what possible drug resistance genes it contains and the specific virulence factors, then the limited resources could be applied in the most effective manner. This thesis will attempt to utilise the increasing availability of WGS and routine sequencing in healthcare institutions to implement an analysis pipeline that might be able to elucidate the possible geographic origin of an MRSA isolate.

1.4 Using the data generated by WGS and NGS techniques

The increasing viability of WGS has enabled investigation into detailed questions in the epidemiology of MRSA. A standard way of analysing epidemiological data, especially when attempting to identify transmission events and the origin of a pathogen, is to construct phylogenetic trees from the data. This methodology shows clustering of highly similar genomes which, when coupled with metadata of sampling location, could give an indication of the geographic origin of an isolate. If it appears that the isolate has originated from a location other than where it was sampled from, then this could be an indication of a transmission event. This process is further developed later in the thesis, particularly in Chapter 3. Many of the studies of MRSA that utilise WGS and NGS create phylogenetic trees of the sequenced data (see Section 1.4.2). However, there are many different methodologies available for the construction of phylogenetic trees, each with their own strengths and weaknesses.

1.4.1 Phylogenetic analysis techniques

Phylogenetics may use morphometric or molecular data to identify the evolution and divergence between species or, as in the case of MRSA, within a species. The goal is to create a phylogenetic tree (i.e. a phylogeny) of the evolution of a group of taxa, genes, or characteristics. A phylogenetic tree consists of a branching structure, where each bifurcation is indicative of an evolutionary split, and each taxon as a separate terminal node. Each dataset has a number of possible phylogenetic trees, defined as the "tree space". These trees may be rooted or unrooted. Rooted trees compare the input sequences to a specified most recent common ancestor (MRCA), while unrooted trees determine the relationships between the input sequences. For sufficiently large datasets, attempting to reconstruct all the phylogenetic trees is computationally prohibitive. Therefore, search paths, using optimisation criteria, can be used to determine the "best" tree. However, the tree identified may only be optimum at the local level and not the global level. For example, in the hill-climbing method, which incrementally changes a single element in the tree in an attempt to find a better solution (Ganapathy *et al*., 2003), it is possible to arrive at a local optimal tree, whereas the Monte Carlo tree search methods fares better at finding the global optimal tree (Browne *et al*., 2012).

Historically, phylogenetics developed using morphological characters, however molecular characters can be more informative in modern phylogenetics (Huson & Bryant, 2006). Molecular characters could be nucleotides in DNA and RNA (i.e. genetic characters, SNPs), amino acids, or distinct gene alleles. These characters are used to generate a measure of genetic dissimilarity, or distance, between any two taxa, or nodes. Genetic distance

measures can create phylogenies with each individual input sequence assigned to a separate terminal node and the distances of the branches proportional to the genetic distances (Huson & Bryant, 2006). However, to determine which characters are homologous requires the alignment of all the sequences in the group, which may be troublesome if there are sufficient mutations, insertions, or deletions (Hogeweg & Hesper, 1984). There are a number of different methods of constructing a phylogeny; some of the most commonly used are Maximum Likelihood, Maximum Parsimony, Bayesian inference, and Neighbour Joining (Kuhner & Felsenstein, 1994). These methods may be parametric and require an underlying mathematical model, or non-parametric and use distance, or dissimilarity, measures (Huson & Bryant, 2006).

The Neighbour Joining (NJ) method is an agglomerative clustering method first proposed by Saitou & Nei (1987). It uses a distance matrix $\mathbf{Q}$, specifying the distance (i.e. dissimilarity) between each pair of nodes. The pair of unique nodes *i* and *j* for which $\mathbf{Q}_{ij}$ has the lowest value are identified. These two nodes are then combined into a newly created node. The distances between this new node and all other nodes are calculated, resulting in a new distance matrix. The method then repeats until all nodes are clustered. The NJ method is rapid and often used for large datasets where other analysis techniques might be computationally prohibitive (Day, 1987), such as large DNA or protein sequences (Didelot, 2010). However, it has been superseded by other methods which do not rely on distance metrics, such as Maximum Likelihood.

Maximum Parsimony (MP), a non-parametric method, identifies the phylogenetic tree which requires the smallest number of evolutionary events to explain the observed data (Farris, 1970; Fitch, 1971). This idea of parsimony is prevalent in most phylogenetic methods ; i.e. a simpler chain of events required to obtain an output is favourable over a more complex one (Jaynes & Bretthorst, 2003). MP analysis works by giving each tree a parsimony score. The tree with the greatest parsimony is the favoured one. Since it would be impractical to exhaustively search for all possible trees for a large dataset, the favoured tree of the previous step is perturbed to see if a more parsimonious tree may be achieved. However, it was shown that MP may be inconsistent in a number of ways; for example long-branch attraction is a systematic error where distantly related taxa are incorrectly inferred to be closely related due to both taxa undergoing large amounts of change (Felsenstein, 1978).

Maximum Likelihood (ML) in phylogenetics was first attempted in 1964 (Edwards & Cavalli-Sforza, 1964), with the first nucleotide-based data attempt in 1974 (Neyman, 1974). It is one of the most popular alternative phylogenetic methods even though it is more

computationally demanding (Guindon & Gascuel, 2003). ML is a parametric statistical method which uses optimality criteria to determine the best topology of the phylogenetic trees. In ML, the tree with the highest maximum likelihood score is preferred (Swofford *et al*., 1996). Furthermore, ML requires an underlying model of the evolution of the characteristics which must be a reasonable approximation of the processes that produced the data. An incorrect model can produce a biased result. The stochastic model used in constructing a phylogeny with ML gives the probability of the change of a particular character. The model can have an exorbitant number of parameters, which could encompass probabilities of particular states, particular changes or differences in change among characters. The ML method usually produces trees which are very similar to the most parsimonious tree for the same dataset, and with fewer of the disadvantages associated with MP (Kolaczkowski & Thornton, 2004).

Bayesian inference phylogenetics is based on the probabilistic method developed by Thomas Bayes (Bayes & Price, 1763). Bayesian inference phylogenetics creates trees by constructing a posterior probability, using a model of evolution, from a likelihood function and prior probabilities, providing the most likely phylogenetic tree for the given data (Gelman *et al*., 2014a). It requires a large amount of computing power and so it has only relatively recently become popular. Furthermore, the large number of possibilities available when constructing the posterior probability results in a huge number of possible trees which require the implementation of Markov Chain Monte Carlo algorithms to sample and obtain the optimum tree (Z. Yang & Rannala, 1997). These algorithms also require substantial computing power. Therefore, Bayesian inference phylogenetics is not often used for large datasets, although it is often believed to be the most accurate of the phylogenetic models available, provided the correct parameters are implemented (Wiens & Moen, 2008).

All parametric phylogenetic methods require a mathematical model that attempts to describe the evolution of the characters in the data. These models often make assumptions, either explicitly or implicitly about the input data. Thus, the output of any phylogenetic analysis is only a hypothesis for the evolution of the data, and may be inaccurate and biased. Furthermore, parametric molecular phylogenetic methods require the use of a defined substitution model referring to the rate of mutations at the various character sites examined (Sullivan & Joyce, 2005). The Jukes-Cantor model is the simplest and assigns an equal probability to all nucleotide bases (Jukes & Cantor, 1969). The models may become progressively more complicated to take into account various aspects of unequal mutation rate; for example, correcting for differences in transition and transversion rates, or correcting for context-dependent evolution of nucleotides (Siepel & Haussler, 2004). The selection of the

most appropriate model for the data is critical in obtaining a relevant phylogeny. Although there are methods to determine which model would best suit a given dataset (e.g. likelihood ratio test (Huelsenbeck & Bull, 1996), Akaike information criteria (Akaike, 1981), or Bayesian information criteria (Schwarz, 1978)) care must still be taken when choosing a model. Therefore, there is still some subjectivity on behalf of the investigator as to which model is chosen and presumed to best fit the data.

Other than violating the assumptions mentioned there are a number of problems that may be encountered when attempting to construct the "true" phylogenetic tree. Firstly, certain characters, especially in molecular phylogenetics, may be the result of homoplasy. That is, two (or more) separate originations of the same character, leading to the erroneous assumption that the organisms in question are related through this character. This could be attributed to convergent evolution or due to random mutations at the same site. Secondly, organisms may inherit genetic material in two ways; vertical gene transfer is the passing of genetic material from parent to offspring, while horizontal gene transfer (HGT) is when genetic material may be passed between unrelated organisms. This process, if unaccounted for, may create artefacts in the phylogenetic reconstruction of a group of organisms. Missing data, lack of taxa and lack of appropriate characters all may play a part in providing a false phylogenetic output. Finally, recombination may, if unaccounted for, lead to an overestimation of the mutation rate heterogeneity or underestimate the timings of events, such as pathogenicity acquisition (Awadalla, 2003).

It is possible to determine how much support to assign to a particular phylogeny. Determining support can be done through a few different methods such as jack-knifing or bootstrapping (Phillips *et al*., 2004). Jack-knifing is a resampling technique which removes a single observation from the data and recalculates the output. This is repeated and the average value of each jack-knife is retained. In this way this resampling technique tests the bias, if any, of the tree. In bootstrapping all the data are resampled and the consistency with the original output is determined. This is repeated, usually between 100 and 500 replicates (Pattengale *et al*, 2010), and the number of consistent bootstraps lends support to the original output.

Phylogenetic analysis in epidemiology is useful in answering questions related to transmission events or population structure. For example, the shape of the phylogeny is an important feature which may elucidate information of the population; star-like phylogenies are indicative of either a recent population expansion or recombination (Awadalla, 2003). However, a phylogenetic approach is only valid with the assumption that recombination does

not occur, or occurs at negligible levels (Awadalla, 2003). If recombination has occurred then methods which explicitly take this into account must be used, though all phylogenetic and comparative methods differ in their ability to detect recombination (Awadalla, 2003; Posada & Crandall, 2001). Since recombination allows genomic regions to have independent evolutionary histories, phylogenetics have been used to determine the extent to which recombination is occurring in a genome, which is essential for locating pathogenicity loci (Awadalla, 2003). Therefore, provided care is taken, phylogenetic analysis is an important tool in the study of epidemiology.

1.4.2 Examples of WGS and NGS in the investigation of MRSA

WGS investigation outperforms traditional methods (e.g. *spa* typing) in identifying transmission events in a number of studies. In one study by Price *et al*., (2014) the *spa* typing method falsely suggested transmissions between patients, and failed to identify other transmission events. This finding was obtained through WGS of the isolates and maximum-likelihood phylogenetics. A further study by Török *et al*. (2014) of five MRSA bacteraemia isolates from a Cambridge University hospital showed that MLST indicated high relatedness between these isolates with four of them from ST22 and one from ST2046. However, WGS demonstrated that the cases were actually all unrelated. This indicates the frequent introduction of MRSA into this healthcare institution, rather than spread within the institution. Another study by Harris *et al*. (2013) in a special care baby unit in Cambridge identified 26 related cases of MRSA ST2371 (belonging to the same CC as EMRSA-15) carriage and used WGS and phylogenetics to show that transmission events occurred both in the healthcare institution and in the community. Furthermore, WGS was able to confirm that a staff member of the healthcare institution carried the ST2371 strain between known infections and therefore allowed the outbreak to persist. These findings could not be resolved with conventional methods. This study shows that WGS allows for rapid identification of subtle differences in an isolate's genome and SNP variation that can be used to trace isolates during outbreaks and determine their relatedness. Köser *et al*. (2012) constructed a phylogenetic tree using SNPs of an EMRSA-15 clone (ST22). They found a distinct clustering of outbreak isolates which were separated from non-outbreak isolates, with the identification of a previously missed transmission event. They also identified a hyper-mutator strain, indicating that a simple SNP cut-off threshold to indicate relevance to transmission events is an invalid approach. Furthermore, they showed that the use of WGS is clinically valid with regards to rapidity of analysis, at no (or negligible) extra cost. This study further highlights the value of WGS in real time control of MRSA in healthcare institutions. However, the authors do note that automated

interpretation of the WGS data is likely to be a necessity to make this a practical approach. Another study, by McAdam *et al*. (2012) on isolates from CC30 which includes the pandemic EMRSA-16 clone, showed molecular correlates of MGEs and non-synonymous mutations affecting virulence and antibiotic resistance with HA and CA pandemics. This study used Bayesian phylogenetic construction to show that EMRSA-16 in the UK spread to smaller healthcare institutions from hospitals in large population centres. This study also implicates patient transfers as the vector by which MRSA spreads, which supports the work by Donker *et al*. (2012; 2014). The McAdam *et al*. (2012) study showed the promise of using WGS and phylogenetic techniques to track the emergence and transmission of an epidemic MRSA clone.

Using WGS Harris *et al*., (2010) revealed the temporal and global spread of multi-drug resistant MRSA strain ST239. They used core genome SNPs found through WGS to create a maximum likelihood phylogeny of the hospital and intercontinental spread of ST239. Little homoplasy was identified and the few homoplasic SNPs were in genes known to be involved in drug resistance. The phylogeny showed a consistent geographic clustering, with the European isolates clustered basally on the tree. This is consistent with an European origin for this clone, with the most recent common ancestor dated to the mid-1960s. However, several exceptions to this clustering indicate inter-continental spread; for example, two European clones clustered within the Thai clade. This study also enabled fine scale transmission events between or within hospitals to be identified; for example, five isolates from a Thai hospital differed only by 14 SNPs. This finding has important implications for targeted infection control, since it is possible that a single SNP can distinguish between highly related isolates. This study highlights the need for more global surveillance strategies.

The use of WGS could help reduce the burden of nosocomial infection in resource-restricted healthcare settings. Tong *et al*. (2015) used WGS techniques to sequence 79 ST239 isolates from a hospital in northeast Thailand and found, using a maximum likelihood phylogenetic analysis, that there was distinct variation in ST239 clades over time. This is indicative of cycles of introduction, transmission and extinction. Furthermore, this study identified variability in the particular resistance encoding genes and MGE complements. Therefore, the information of multiple transmission and introduction events obtained through the use of WGS would enable the targeting of the limited resources available.

WGS has also been used to investigate the zoonotic transmission to and from animal reservoirs; for example, Price *et al*. (2012), Harrison *et al*. (2013; 2014), and Spoor *et al*. (2012). Price *et al*. (2013) investigated the livestock associated clonal complex CC398. They

obtained 4238 SNPs through WGS of 89 core genomes and used this information to construct a maximum likelihood phylogenetic tree. The tree strongly suggests an MSSA human origin for CC398, with rapid radiation with the first human to animal transmission and the subsequent acquisition of tetracycline and methicillin resistance. Furthermore, the diversity of SCC*mec* subtypes in this CC is suggestive of antimicrobial selection pressures. In the Harrison *et al*. (2013) study isolates (both human and animal) were taken from two Danish farms associated with LA-MRSA infections. The isolates from the two farms were indistinguishable by conventional techniques (e.g. MLST, PFGE, *spa*-typing). However, phylogenetic analysis of the WGS data showed two distinct farm-specific clusters of isolates. The isolates sampled from the humans and livestock of the same farm only differed by a few SNPs, indicating a zoonotic origin. The study by Spoor *et al.* (2013) showed, using the mainly bovine complex CC97 and high-resolution phylogenetics, the limited number of evolutionary events required for a zoonotic jump to occur. The authors concluded that livestock represent a reservoir of MRSA with major public health implications. It is not just livestock which may be reservoirs for MRSA. Harrison *et al*. (2014) investigated the ST22 strain in companion animals, such as cats, horses and dogs. The authors found evidence for a human source of the isolates infecting companion animals. The studies mentioned here do raise some concerns as to the validity of the "one health" view of infectious diseases; that is that the pathogens which show zoonotic capabilities intrinsically link multiple species. This implies that antibiotic use in non-human species might lead to antibiotic resistance in a pathogen which might then make the jump into humans. This complex zoonotic epidemiology will be difficult to untangle, and the use of WGS and appropriate tracking techniques will be important in combating the spread.

Furthermore, the use of WGS and phylogenetics in a study by Paterson *et al*. (2015) illustrates the considerable within-host diversity and fluctuation in the diversity in both human and animal patients. Therefore, this study demonstrates the need for the sequencing of multiple isolates from individuals to elucidate the accurate transmission networks. Computer simulations conducted by Worby *et al*. (2014) show that sequencing a single isolate from each host is inadequate to obtain transmission networks and may lead to misleading interpretations. The authors further conclude that the use of sequence data alone is not sufficient, and other traditional methodologies (e.g. identification of overlapping admittance to a healthcare institution) are required. Finally, work by Colijn & Gardy (2014) showed that it is possible to characterise transmission events based on simple topological properties of the phylogenetic tree constructed from WGS genome data.

These examples clearly show the advantages of WGS compared to traditional methodologies, especially in the study of transmission routes and timing events of MRSA. However, most of these examples provided have used relatively small datasets to generate the phylogenetic trees. If WGS becomes a routine procedure in healthcare institutions with a corresponding online database, then there will soon be too many isolates to realistically conduct phylogenetic analysis, since this approach may become cumbersome and impractical for large datasets. Furthermore, as was noted by Köser *et al* (2012), there is a need for automation to make this financially and practically viable. Therefore, there is an opportunity here to develop novel methods that could identify the origin of an MRSA isolate automatically, without resorting to phylogenetic construction.

1.5 Conclusion and summary of thesis

The advancements in Whole Genome Sequencing have paved the way for the development of novel analysis techniques. WGS has allowed for the generation of large datasets, and has already provided crucial insights into the epidemiology of MRSA. However, phylogenetic analysis of large datasets is computationally prohibitive, especially if attempting to combat an ongoing outbreak by rapidly answering such epidemiological questions as the origin and transmission of a MRSA isolate. Therefore, this thesis presents the development of novel methods which will attempt to elucidate the possible geographic origin of a given MRSA isolate, within the confines of the sampling dataset. These novel methods will attempt to remove the necessity of the construction of a phylogenetic tree. However, these methods would be applicable to isolates within a clonal complex, since alternative methodologies (such as MLST or PFGE) are robust at defining isolates up to this resolution. Therefore, the application of these methods would fall within an analysis pipeline that would include conventional genotyping techniques. Each chapter is built on the findings of the previous one.

The primary dataset used in this thesis is a collated ST22 MRSA dataset comprising 1022 isolates sampled from 46 hospital locations across the UK and Ireland between 2001 and 2010. However, there are some isolates in the collated dataset sampled in 2011 and 2012, which are used in Chapter 4. Furthermore, isolates obtained from Holden *et al* (2013) are also used in Chapter 4. Finally, I used an indel dataset in Chapter 6. These additional datasets are described in their respective chapters. The primary dataset is examined in detail in Chapter 2 at two geographic resolutions; either with individual hospitals or with a group of hospitals (defined as a Referral Cluster) as the MRSA sub-populations. The SNPs found in the isolates are used to draw phylogenetic trees and it is found that there is phylogeographic clustering. Further examination shows that the SNPs are in fewer locations than expected for the number of isolates they are harboured in. The SNP similarity of the MRSA sub-populations is influenced by the geographic proximity of the locations, and the number of patient transfers between locations. Therefore, this shows that there is structure in the MRSA sub-populations, and the movement of strains from one location to another is important in maintaining the genetic heterogeneity of the sub-populations.

Chapter 3 explores the identification of introduction events of MRSA from one location to another. Introduction events were determined using a phylogenetic approach. Each isolate defined as an introduction was examined to identify if any of the SNPs harboured in that isolate could be defined as a signature SNP for a particular location. A signature SNP is one that

is only ever seen in one location, and therefore could be a good indicator that the introduction may have originated from that location. It was found that a select few of the introduction events can be characterised by a signature SNP from the posited origin location. Therefore, it may be possible to identify introduction events by looking at the SNPs an isolate contains.

In Chapter 4 this idea is developed further. A novel method is developed where the SNPs harboured in an isolate are examined for geographic signal, and converted into a diagnostic value of origination for each of the locations in the dataset. This method is termed the SNP-based Assignment of Pathogen Origin (SnAPO). The isolates identified as introduction events in Chapter 3 were first used to test SnAPO, since it might be expected that there would be some signal indicating that the isolate is a transmission event. SnAPO was then tested using all isolates sampled in 2010 as test cases, where there was not necessarily an expected introduction event signal. It was found that SnAPO is able to give a posited origin location of any isolate, although there is great variation in the signal clarity. The output of SnAPO was then compared to the results obtained by three independent researchers who used a phylogenetic approach to determine the possible origin of the isolate, with the majority of test cases conforming between the phylogenetic approach and SnAPO. However, the subjectivity of the phylogenetic approach was seen in the variation of the posited origin locations for the test isolates, while SnAPO is consistent and objective. Further testing of SnAPO was conducted on BSAC isolates sampled in 2011 and 2012 which had their location metadata removed. Finally, an alternative dataset, extracted from Holden *et al* (2013), was used to show that SnAPO could work on isolates from a different dataset. Therefore, a novel method has been developed which is objective, fast, simple and obviates the phylogenetic requirement.

Chapter 5 explores the development of an alternative approach to determine the possible geographic origin of an isolate using a Bayesian inference approach. This would use established statistical approaches and move away from the heuristic SnAPO method. It was found that the Bayesian approach concurs with SnAPO for the origin location in the majority of test cases.

Finally, Chapter 6 explores some of the possible limitations of SnAPO and the modification of SnAPO for an indel dataset. The robustness of the method was investigated by removing isolates from the dataset and it was found that SnAPO is robust to changes in the dataset since it predicts the same origin location in the majority of the test isolates. It was found that older isolates may be obscuring some information and so an optimum dataset size of 6 years prior to target isolate was identified. The SnAPO method was applied to an indel

dataset and it is found that the indel data can also be used to generate a posited origin location for an isolate that matches the one obtained from the SNP data. Therefore this method may be applicable to other systems. Finally, the effect of degrading the signal expressed in a focal isolate's SNPs was explored, and it was found that an isolate could have approximately a third of its SNPs replaced before a different origin location is posited.

In this thesis I present the development of novel methods which obviate the need for a phylogenetic tree and determine the possible geographic origin of an MRSA isolate within CC22. These methods are rapid, objective and easily interpretable. These properties will be useful to quickly determine if an isolate is an introduction, and if so, potentially warn of the beginning of an epidemic spread while it happens. In this way the limited resources available may be better focussed to combat the spread. Furthermore, I show that the principles developed in this thesis may allow for deeper understanding of the rich diversity of information available through WGS, and enable detailed investigation into the transmission and evolution of pathogenic bacteria.

2 Characterisation of the dataset

2.1 Background

Methicillin-resistance *Staphylococcus aureus* (MRSA) was first isolated in the UK in 1961 (Jevons, 1961). The prevalence of MRSA in the UK gradually increased, with full UK coverage observed by the mid-1980s (Johnson *et al*., 2005). However, mandatory reporting of all cases of *S. aureus* bacteraemia and the number due to MRSA only began in 2001 (Pearson *et al*., 2009). Prior to 2001 voluntary reporting was the routine method of surveillance. Mandatory reporting, coupled with an increase in active surveillance, has allowed the development of larger MRSA genomic collections.

There are a number of collections of MRSA genomes in the UK. One collection, numbering several thousand genomes, is the British Society of Antimicrobial Chemotherapy (BSAC), which coordinates an antimicrobial resistance surveillance project (www.bsacsurv.org; Reynolds *et al*., 2008). Each hospital or laboratory is required to submit the first 10-14 clinically significant *S. aureus* genomes each calendar year to the BSAC collection. The collection is further bolstered in size by voluntary MRSA isolate submissions. This MRSA collection is one of the largest of its kind in the world; the product of many different collaborating laboratories and scientists. This collection provides unique opportunities for investigation and is a powerful resource for the analysis of MRSA outbreaks.

One of the main aims of the BSAC collection was to understand the population structure and dynamics of MRSA spread across UK and Republic of Ireland over the last decade. Reuter *et al*., (2015) found that the majority of isolates belonged to Clonal Complex (CC) 22, which contains the dominant UK epidemic clone (EMRSA-15). The second most frequent CC in the UK was CC30, containing EMRSA-16. Isolates from a number of other CCs comprise the rest of the collection. Geographic structuring of EMRSA-15 was found to be consistent with widespread dissemination followed by local diversification prior to the sampling period. Local and regional differences in antibiotic resistance were also observed. Finally, research on an England-only subset of the BSAC collection by Donker *et al.* (2012, 2014) showed that there is phylogeographic clustering of the isolates based on the healthcare

Referral Cluster (RC) they were sampled from. An RC is defined by the level of patient transfer, with higher numbers of patient referrals within than between an RC.

The central contention of this thesis is that it might be possible to obtain a clearer picture of how pathogenic microbes, in this case MRSA, might evolve and transmit by designing analytic tools which focus on genetic variation SNP-by-SNP. The goal of this chapter is to assess the extent to which SNPs spatially and phylogenetically cluster. In this chapter I describe the collated single nucleotide polymorphism (SNP) dataset of 1000+ MRSA genomes taken over a decade from across England, Scotland, Wales, Northern Ireland and Republic of Ireland (UK&I). Isolates sampled in the same geographic location could be considered to be from the same MRSA sub-population. These 1000+ isolates are all CC22 isolates and are taken mainly from the BSAC collection, with some taken from the East of England (EoE) collection. The EoE collection mainly contains isolates sampled in Cambridge Addenbrooke's Hospital and Papworth Hospital. I investigate the SNPs to identify similarity between isolates and MRSA sub-populations. The traditional phylogenetic approach indicated phylogeographic clustering of genetically similar isolates. This was further supported by the finding that some SNPs are in fewer locations than expected for their relative abundance. Finally, I used the level of SNP similarity between isolates and MRSA sub-populations to show that there is differentiation in the MRSA genetic structure of the sub-populations, which may be attributable to the level of patient referral and geographic proximity of the sub-populations.

2.2 Methods for obtaining the data

The genetic information of the 1022 isolates used in this thesis was determined as described in Reuter et al. (2015) which is summarised here. DNA extraction was conducted on a QIAxtractor (QIAGEN) and library preparation was performed as described in Köser et al. (2012). Index-tagged libraries were created, and sequenced using the Illumina HiSeq platform (Illumina Inc.) to generate pair ended reads of 100 base-pairs (bp) at the Wellcome Trust Sanger Institute. The pair-end reads were mapped against the chromosome of *S. aureus* HO 50960412 (accession no. HE681097). Indels were identified using Dindel (Albers et al., 2011). SNPs were identified using ssaha_pileup, with filtering to remove those sites with a SNP quality score less than 30 and those SNPs at sites with heterogeneous mappings, if the SNP was present in less than 75% of reads at that site (supplementary material, Harris et al., 2010). SNPs from unmapped reads or sequences that were not present in all genomes were not considered part of the core genome and therefore excluded from the analysis. Furthermore, SNPs falling within MGEs regions, and those within high density regions, were also excluded. Additionally, the SCC*mec* element, which contains the *mecA* antimicrobial resistance gene, was removed prior to receiving the data from BSAC and EoE. The core genome was curated manually for increased quality assurance and is comprised of 2,643,131 bp. The SNPs in the core genome were extracted, reducing down to the 29651 SNPs used for analysis in this thesis, both phylogenetic and otherwise. Therefore, the data collated from the BSAC and EoE collection and used in this thesis contained 1022 isolates with 29651 SNPs (see Section 2.4 for description of the SNPs). This will be termed the Unmodified Dataset.

2.3 Isolate sampling

All 1022 isolates in the Unmodified Dataset were sampled between 2001 and 2010 in 46 hospitals across England, Scotland, Wales, Northern Ireland and the Republic of Ireland (UK&I). The number of hospitals sampled each year varied slightly over the decade (median = 25, range = 23 – 32). Using a different collated dataset from the BSAC collection, which included a large number of hospitals located in England, Donker *et al.* (2012, 2014) created geographic regions within which there was a higher level of patient referral than between regions. Each region is termed a Referral Cluster (RC) and it is posited that the movement of MRSA-infected patients is one of the main processes by which MRSA is spread to various locations. If this is true then there may be variation in the genetic diversity of the MRSA sub-populations between different RCs and hospitals. The non-English hospitals have been subsequently grouped into their respective countries as individual RCs (Figure 2.1). However, it

is possible that with increased sampling of hospitals in the other countries of the UK and Ireland, there may be further division into smaller RCs. The sampling was not uniform between all hospitals or years. Some hospitals, and therefore RCs, had much greater sampling than others (Figure 2.2).

Two resolutions of geographic sub-populations were considered in this thesis; the hospital resolution (where each MRSA sub-population contains the isolates sampled in that hospital), and the RC resolution (where each MRSA sub-population contains the isolates sampled in that RC). It must be noted that in most cities, except for Glasgow and London, there are only isolates from one hospital. Therefore, for convenience, the hospital is called by the name of the city. The full name of each hospital in this thesis is provided in Appendix A (Supplementary Table A1).

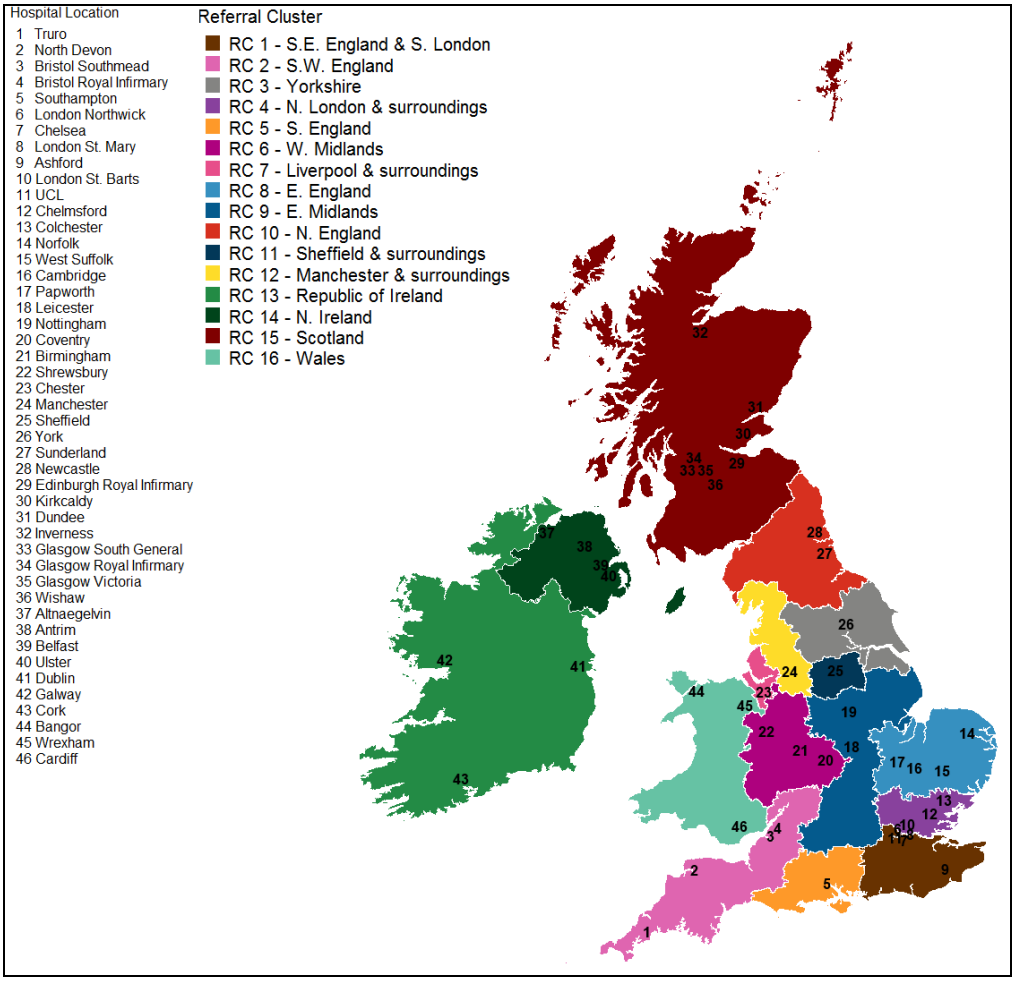


Figure 2.1. A map showing the 46 hospitals present in this thesis, across the United Kingdom, Northern Ireland and the Republic of Ireland. The hospitals have been grouped into 16 separate Referral Clusters (RCs) based on the regions identified in Donker *et al.* (2012, 2014). The hospitals are numbered according to geographic location and RC, with hospitals located in the same RC grouped together.

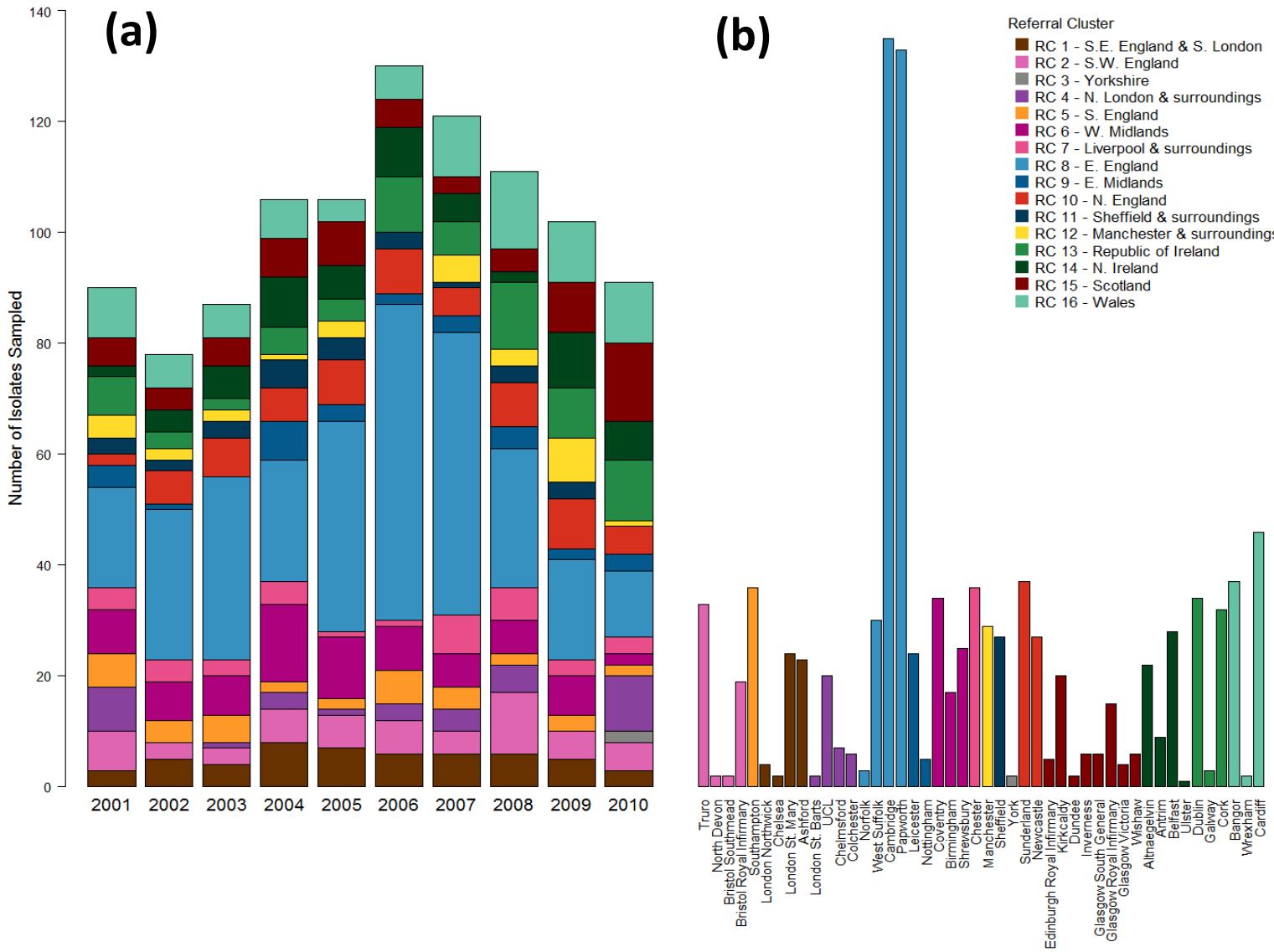


Figure 2.2. The sampling effort in the Unmodified Dataset is broken down by year (a) and hospital (b). In (a) the RC origin of the samples is indicated by the coloured bands. In (b) the hospitals along the x-axis are ordered in the same manner as in Figure 2.1, with hospitals in the same RC grouped together. The key to RC colour is provided in the top right corner of (b). Although there is similar sampling effort across the 10 years (ranging from 78 to 130 isolates per year), there is considerable variation in the sampling across the hospitals (ranging from 1 to 135 isolates per hospital).

47

2.4 Single nucleotide polymorphisms

Phylogenetic tree analysis is often integral in epidemiological studies when determining the possible geographic origin of the isolates (e.g. Harris *et al*., 2013; Köser *et al*., 2012; Price *et al*., 2014), since phylogenetic analysis will group those isolates with high genetic similarity together regardless of sampling location. Phylogeographic structuring would be indicated by phylogenetic clustering of isolates from the same geographic location. Furthermore, transmission events may be identified by the phylogenetic clustering of isolates from disparate locations (e.g. an isolate from Location A is phylogenetically clustered with a group of isolates from Location B, indicating a transmission event from B to A). Therefore there are three terms that will be used throughout this thesis; the "sampling location" is where the isolate was sampled, the "origin location" is where the isolate is thought to have come from, and if the sampling location and the origin location are different then this could be a "transmission event".

In this section I take the traditional initial step of creating phylogenetic trees of the isolates using the SNP data of the 1022 isolates. These isolates are all within Clonal Complex 22, and so depict the evolution within this CC. which may result in isolates which differ by only one SNP. In MRSA this phylogenetic approach is viable since there is a low recombination rate and SNPs are stably inherited (Thomas *et al*., 2011). Therefore, isolates which contain the same SNP could be considered to have shared ancestry. The non-singleton SNPs (i.e. the SNPs harboured in more than one isolate) are then examined to show that there is a limiting factor on the geographic range they occupy. Finally I explore the possible complications arising from the fact that some SNPs are linked.

2.4.1 Constructing phylogenetic trees of the 1022 isolates

Genetic phylogenetic trees may be constructed using full genome or SNP data and are a common way of determining similarity between genomes (Lee *et al*, 2014). There are a number of phylogenetic methods that can use SNP data (discussed in Section 1.4.1), but in this thesis I will use the Neighbour Joining (NJ; Saitou & Nei, 1987) and Maximum Likelihood methods (ML; Siepel & Haussler, 2004). NJ is a non-parametric method and uses a dissimilarity matrix to infer distances between taxa, whereas ML is a parametric method and requires an underlying model of the evolution and nucleotide substitution rates of the genomes. NJ and ML methods are two of the most common techniques used to generate phylogenetic trees, although they require greatly differing levels of computational power. The variation in computational power required is due to the differing trade-offs between speed and accuracy;

the NJ method is fast yet can be inaccurate, while the ML method is slow yet more likely to generate an accurate phylogeny (Kuhner & Felsenstein, 1994). These two methods were chosen since they could result in two different phylogenetic trees.

Two unrooted phylogenetic trees were constructed of the 1022 isolates sampled between 2001 and 2010. One tree was constructed in MEGA5 (Tamura *et al*., 2011) using the NJ method (Figure 2.3). The other was constructed in RAxML (Stamatakis, 2014) using the ML method with a generalised time reduction CAT evolutionary model (Figure 2.4). Both trees were bootstrapped with 1000 replicates and were visualised using the online tool: Interactive Tree of Life (iToL; Letunic & Bork, 2006). The singleton SNPs (i.e. the SNPs harboured in only one isolate) were retained and those known to be associated with drug resistance were excluded in order to reduce the occurrence of homoplasy. As described in Holden *et al* (2013) the literature was mined for the identification of the SNPs associated with drug resistances. The original papers are supplied here for completeness: Aubry-Damon et al., 1998; Castanheira et al., 2010; Chen et al., 2010; Cui et al., 2010; Cui et al., 2009; Griggs, 2003; Gu et al., 2013; Howe et al., 2003; Hurdle et al., 2004; Lannergard et al., 2008; Livermore et al., 2009; Livermore et al., 2007; Locke et al., 2009; Meka et al., 2004; Neoh et al., 2008; North et al., 2005; Prunier et al., 2003; Roberts, 2008; Swaney et al., 1998; Tsiodras et al., 2001; Vester & Douthwaite, 2001; Vickers et al., 2009; Yang et al., 2010. Sampling location information at both the RC and the hospital geographic resolution level was retained. This provided 29627 SNPs and 1022 isolates available with which to construct the phylogenetic trees. This will be termed the Prime Dataset.

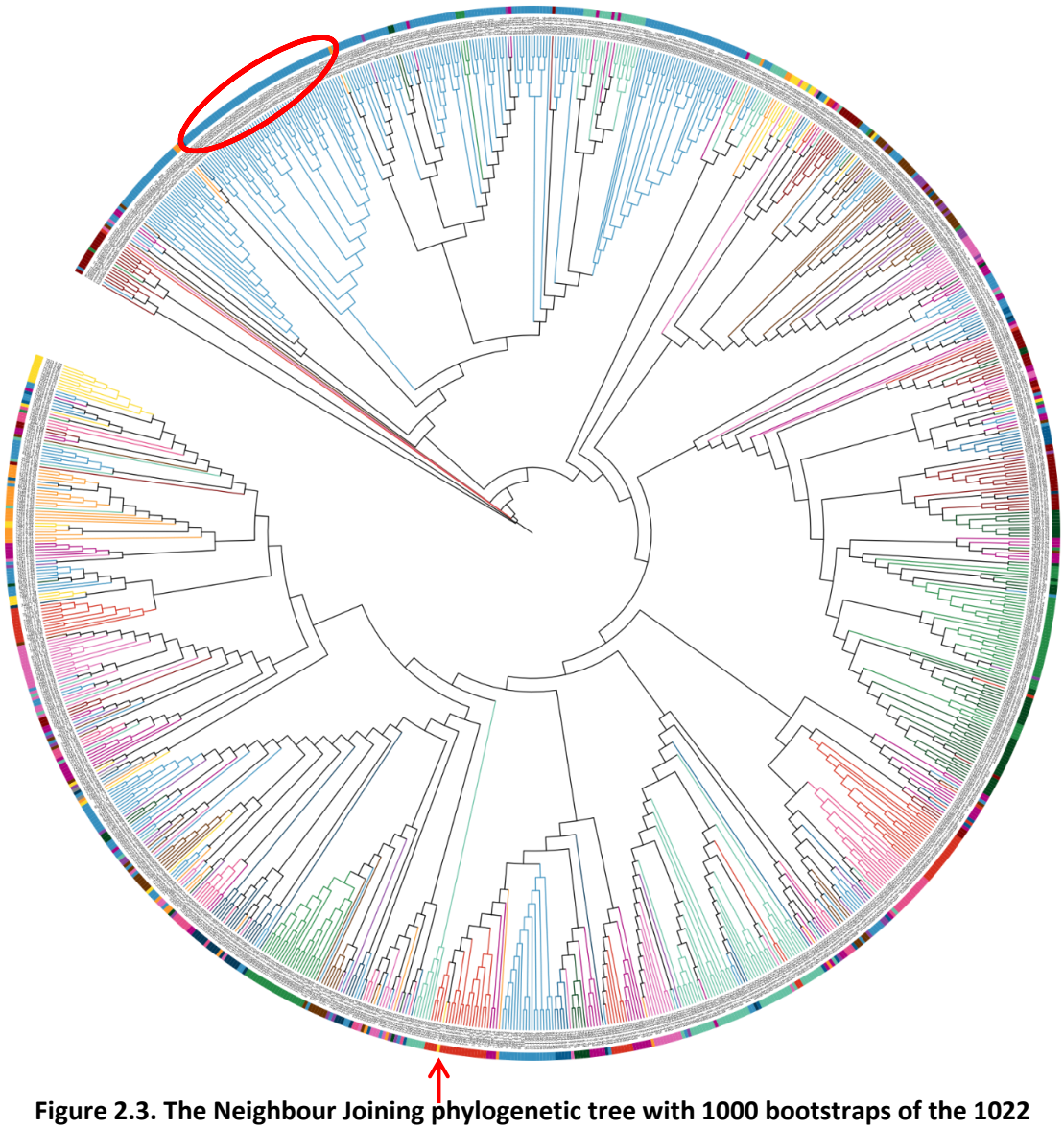


Figure 2.3. The Neighbour Joining phylogenetic tree with 1000 bootstraps of the 1022 isolates sampled between 2001 and 2010 across 46 UK and Ireland hospitals. The taxa labels have been aligned and branch lengths have been homogenised to facilitate visualisation. The colours correspond to the Referral Cluster (RC) each isolate was sampled in, as noted in Figure 2.1. There are some phylogenetic sub-clades with isolates all from one RC (e.g. red circle), indicating there is geographic clustering of genetically similar isolates. However, there are a number of isolates which are located close to isolates from a different RC (e.g. red arrow), indicating a possible transmission event.

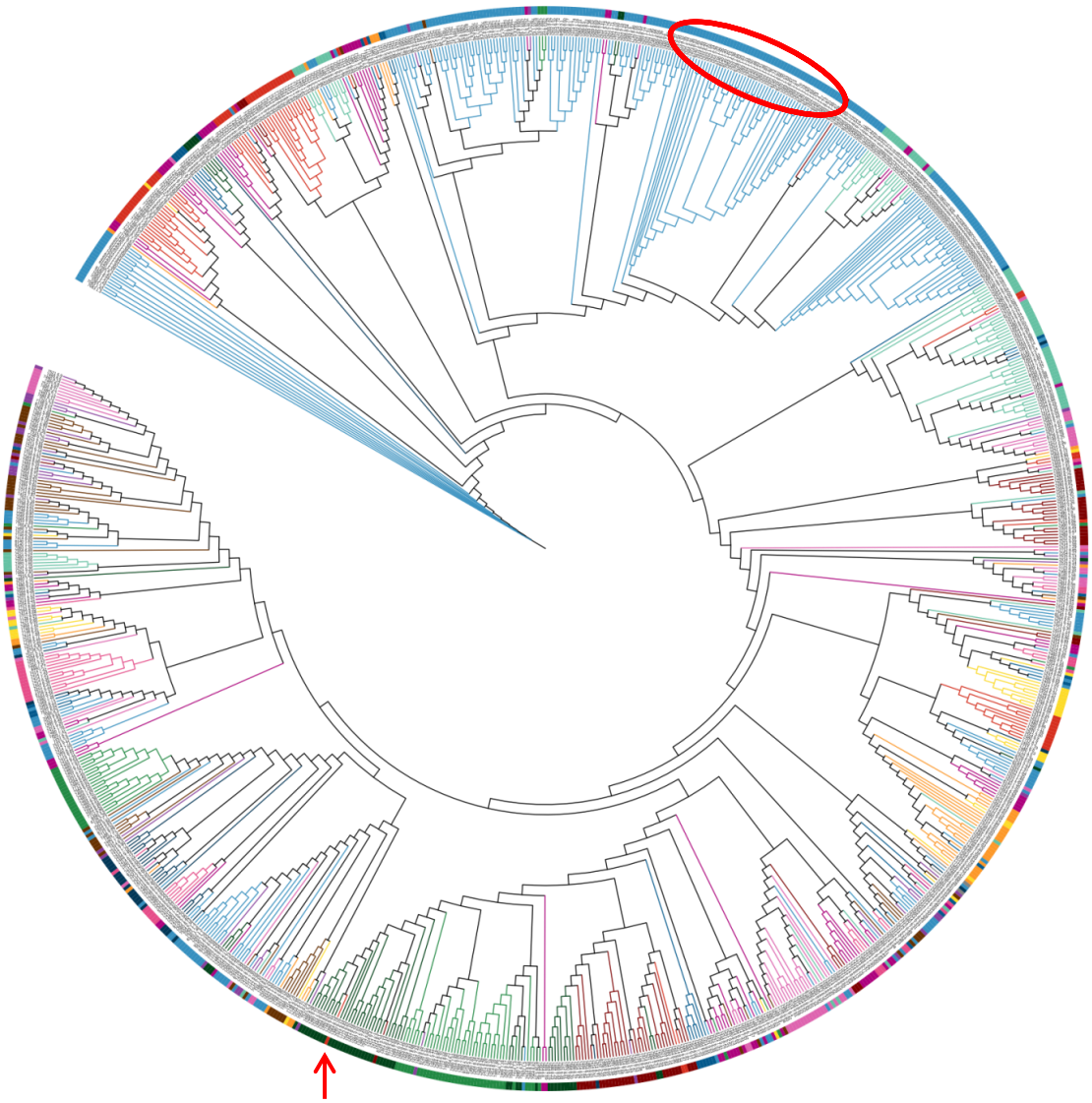


Figure 2.4. The Maximum Likelihood phylogenetic tree with 1000 bootstraps of the 1022 isolates sampled between 2001 and 2010 across 46 UK and Ireland hospitals. The taxa labels have been aligned and branch lengths homogenised to facilitate visualisation. The colours correspond to the Referral Cluster (RC) each isolate was sampled in, as noted in Figure 2.1. As with the NJ tree, there are some phylogenetic sub-clades with isolates all from one RC (e.g. red circle), indicating there is geographic clustering of genetically similar isolates. Similarly to the NJ tree, there are a number of isolates which are located close to isolates from a different RC (e.g. red arrow), indicating a possible transmission event. However, there are a few differences between the two trees, which is developed in this section.

The phylogenetic tree can be manually divided into sub-clades, and isolates within a sub-clade can be considered phylogenetic neighbours and share genetic similarity. In this thesis the sole considerations of sub-clade definition were to maximise the bootstrap and branch length values; the geographic or temporal sampling information was not considered. In certain cases some isolates could not be assigned to a sub-clade. This usually occurred where adding the isolate to the sub-clade would drastically decrease the bootstrap value of that sub-

clade, or where the tree presented a polytomy (see Figure 2.5 for an example). However, these measures to try and standardise the sub-clade definition are not infallible, and so sub-clade definition remains a subjective and time-consuming process.

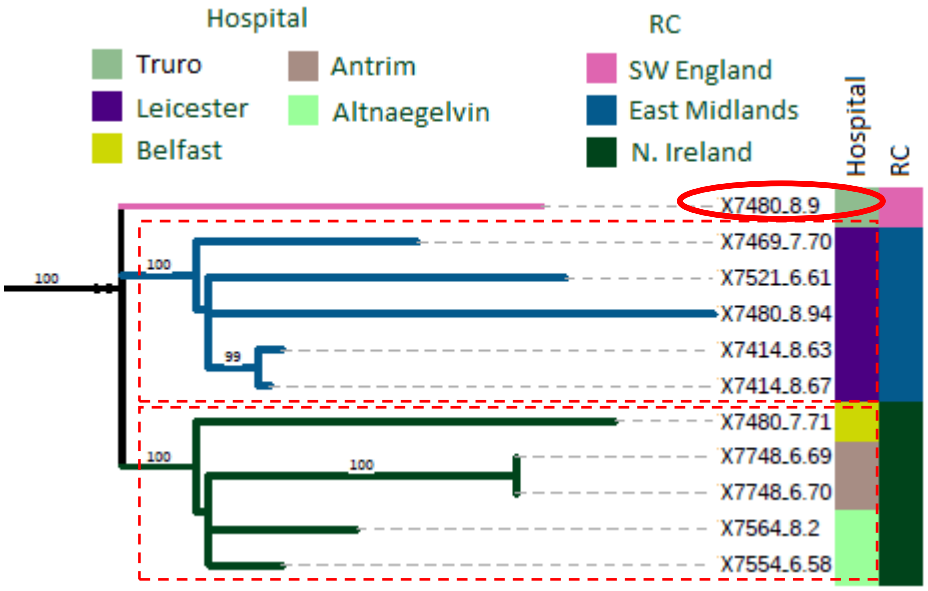


Figure 2.5. Part of the Maximum Likelihood (ML) phylogenetic tree showing two sub-clades (denoted by the red dashed boxes) and an isolate which cannot be assigned to a sub-clade (indicated by a red circle). Branch length is proportional to difference in the number of SNPs in the isolates. The numbers above the branches leading to bifurcations are the bootstrap values. If these numbers are absent then this split has a bootstrap value less than 80. Therefore, in this example this part of the ML phylogenetic tree was split into two sub-clades which have bootstrap values of 100 and clear branch lengths. The branches are coloured by the Referral Cluster where the isolate was sampled, and this colour code is repeated in the right colour column. The left colour column indicates the hospital where the isolate was sampled. The isolate code are used as taxa labels.

There is a range of phylogenetic clustering with regards to geographic location. The sub-clades may contain isolates all from one location, all isolates from a different location, and all variations in between (Figure 2.6). Finally, the consistency of sub-clade clustering between the two phylogenetic trees was determined by displaying them graphically in a Tanglegram (Huson & Scornavacca, 2012; Figure 2.7).

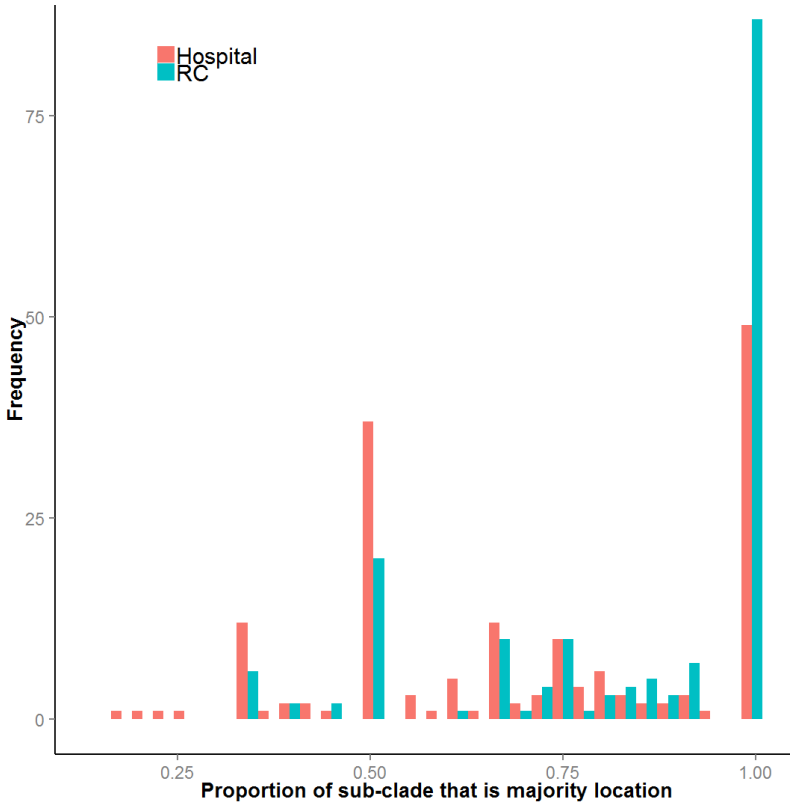


Figure 2.6. Each sub-clade usually contains one geographic location which is more prevalent than the others. The x-axis denotes the proportion of isolates in a sub-clade which were sampled in the most common geographic location for that sub-clade. The majority of sub-clades show geographic grouping at the hospital and RC geographic resolution, with many sub-clades being solely comprised of isolates from the same location. There is a large number of sub-clades which are divided equally between two locations; these sub-clades are likely to be ones with only two isolates. Some sub-clades have a very low proportion of the majority location, indicating that the sub-clade is very geographically mixed with no one location as the clear majority. For identification of transmission events, the most illuminating sub-clades are the ones which contain a high proportion of the majority location (e.g. 90% one location and above).

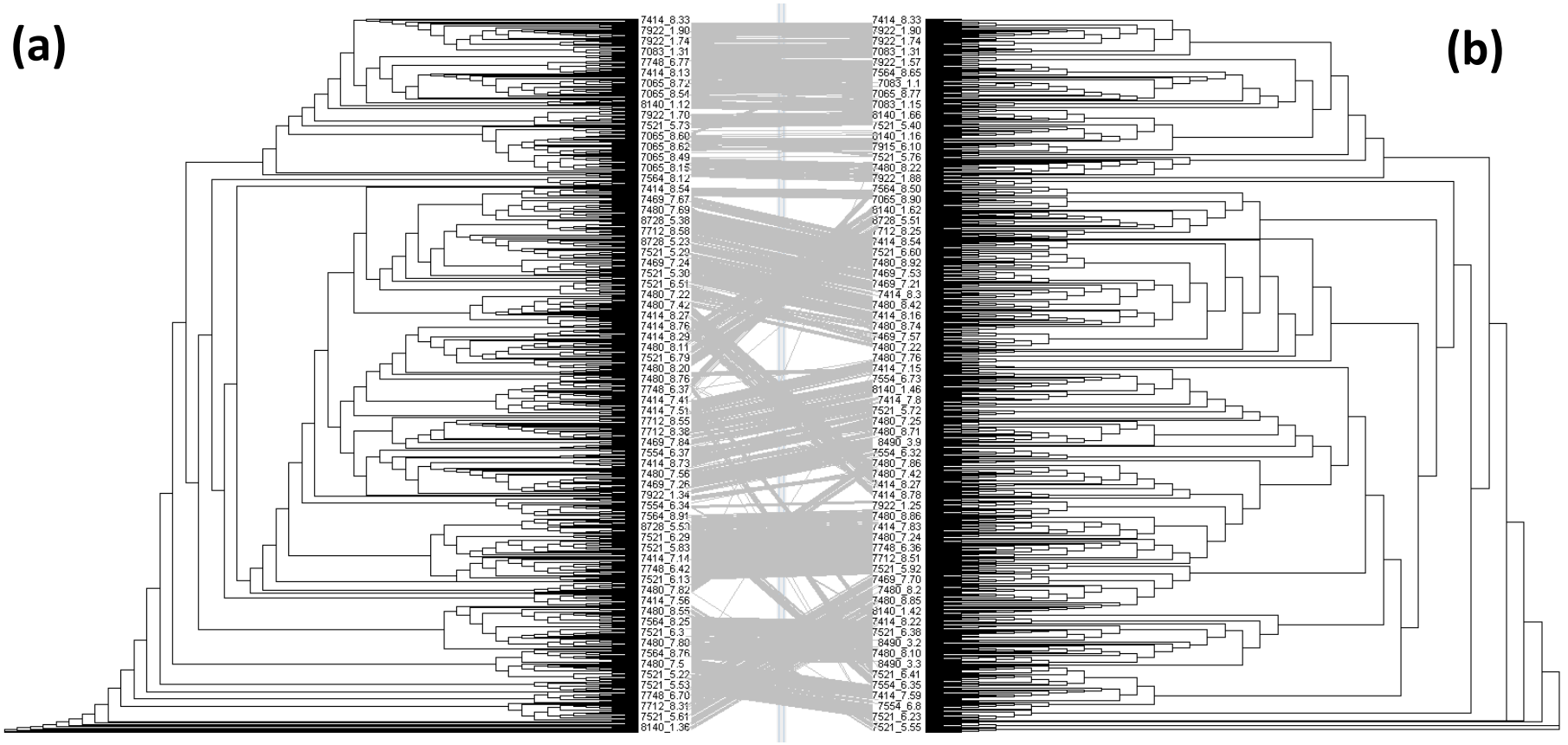


Figure 2.7. A Tanglegram created in Dendroscope showing the differences in clustering between the Maximum Likelihood tree (a) and the Neighbour Joining tree (b). Each isolate is matched up to itself in the opposite tree, creating a number of grey lines. Although all connecting lines are shown, due to the large number of isolates only a few (n = 68) of the 1022 possible leaf labels are displayed. If the two trees were dissimilar there would be a large number of intersecting lines. However, in this example there are a large number of horizontal and parallel lines, indicating that the sub-clade clustering in the two trees is consistent.

The Tanglegram (Figure 2.7) displays a substantial number of horizontal and parallel lines between the two trees. This indicates a high sub-clade clustering similarity, and so the two trees cluster the isolates in a similar way. Although the two methods are creating similar sub-clades it is necessary to ensure that the subclades have similar support in each phylogeny. This can be done by examining the distribution of bootstrap values from each phylogeny (Figure 2.8).

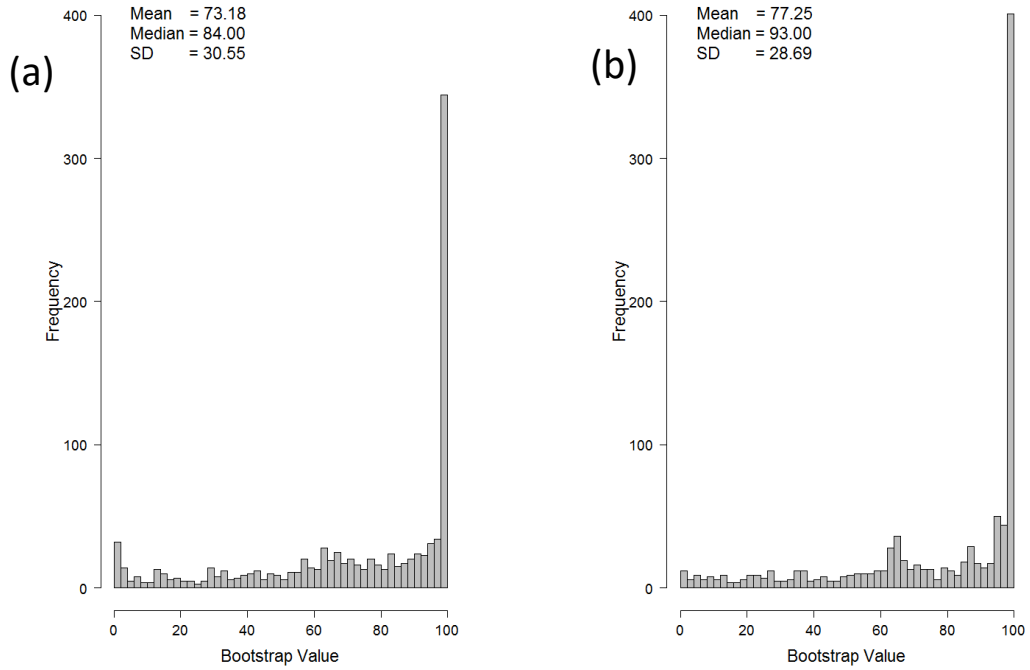


Figure 2.8. The bootstrap values' distributions from the Neighbour Joining (a) and the Maximum Likelihood (b) phylogenies. Both trees used 1000 bootstrap replicates. Although both trees appear to have a large number of well-supported phylogenetic splits, there appears to be greater support in the ML tree, with more splits that are fully supported. This indicates that the ML tree might be a more accurate tree of the evolution of the 1022 isolates within CC22.

The distribution of bootstrap values appear to show greater support of the phylogenetic splits in the ML phylogeny compared to the NJ phylogeny. Therefore, there is no clear distinction between the sub-clade clustering (as seen in Figure 2.7), yet there is a difference in the bootstrap values (as seen in Figure 2.8). Henceforth the ML tree will be used, since it might be a more accurate depiction of the evolution of the isolates within CC22. However, it should be noted that the bootstrap values from the NJ tree still show high support.

2.4.2 Investigating the non-singleton SNPs

Over 80% of the 29627 SNP positions are singletons; i.e. they are only ever found in one isolate. Although useful in the differentiation of isolates from each another (for example,

by increasing branch lengths on phylogenetic trees), singletons are uninformative when comparing similarity between isolates. Therefore these singleton SNPs are removed from the Prime Dataset. Furthermore, those SNPs which are poly-allelic were removed, leaving only bi-allelic SNPs. Some SNP positions have *N*-nucleotides; i.e. those nucleotides which are unknown due to either insertion, deletion or misreading when compared to the reference genome. Those SNP positions which have less than 1% of the 1022 isolates as *N*-nucleotides were retained while the rest were discarded. A threshold of 1% was chosen since those SNP positions with a higher *N*-incidence are likely to be more unreliable. The ones retained had the *N*-nucleotides converted to the majority nucleotide of the SNP position, as the more conservative solution. This resulted in 5469 bi-allelic non-singleton SNP positions spread over the entirety of the circular genome (Figure 2.9). This is termed the Bi-allelic Dataset.

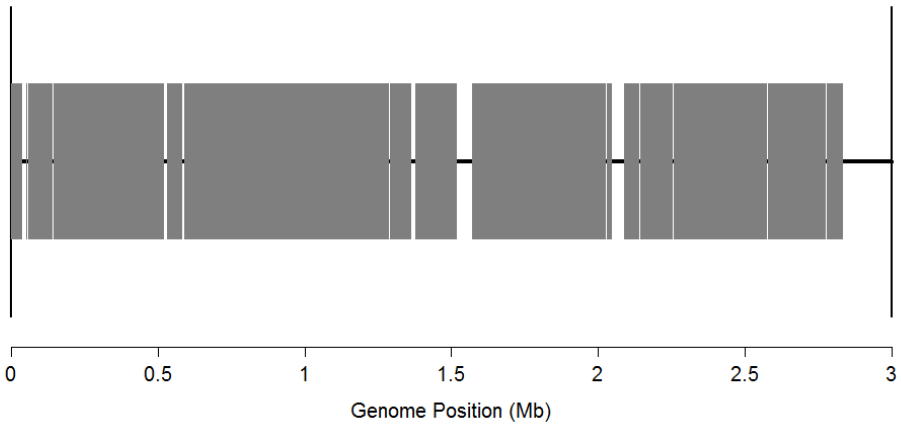


Figure 2.9. A linear diagrammatic representation of the circular genome of approximately 3 million base-pairs of MRSA, with the position of all 5649 bi-allelic non-singleton SNPs. These SNPs are spread out across the whole genome. The SCC*mec* element, which contains the *mecA* gene coding for methicillin resistance and can be spread through horizontal transfer as a plasmid, has been removed from the genome in the Bi-allelic Dataset. The genome position numbering starts from the origin of replication.

At any given bi-allelic SNP position one nucleotide is less common than the other across the 1022 isolates in the Bi-allelic Dataset. This is termed the Minor Allele Frequency (MAF). Therefore, each SNP position can be reduced to a majority and minority nucleotide (Figure 2.10). The majority nucleotide can be considered the genetic background for that particular SNP position, while the minority nucleotide is the mutation that causes that particular position on the genome to be a SNP. Henceforth, the minority nucleotide will be termed as the "SNP", with "SNP position" referring to the position on the approximately 3 million base-pair genome.

It is important to mention that the majority and minority nucleotides might be different in a separate dataset and the particular MAF nucleotide may be different in separate populations. Furthermore, adding more isolates to the Bi-allelic Dataset may cause some of the minority nucleotides to become majority nucleotides. This will only be an issue for those SNP positions where there is close to a 50% incidence of minority nucleotides. In the Bi-allelic Dataset there are only 8 SNP positions which are above 40% minority nucleotide, therefore it is unlikely that the addition of more isolates would cause significant deviation from the identified minor and major nucleotide for any given SNP position.

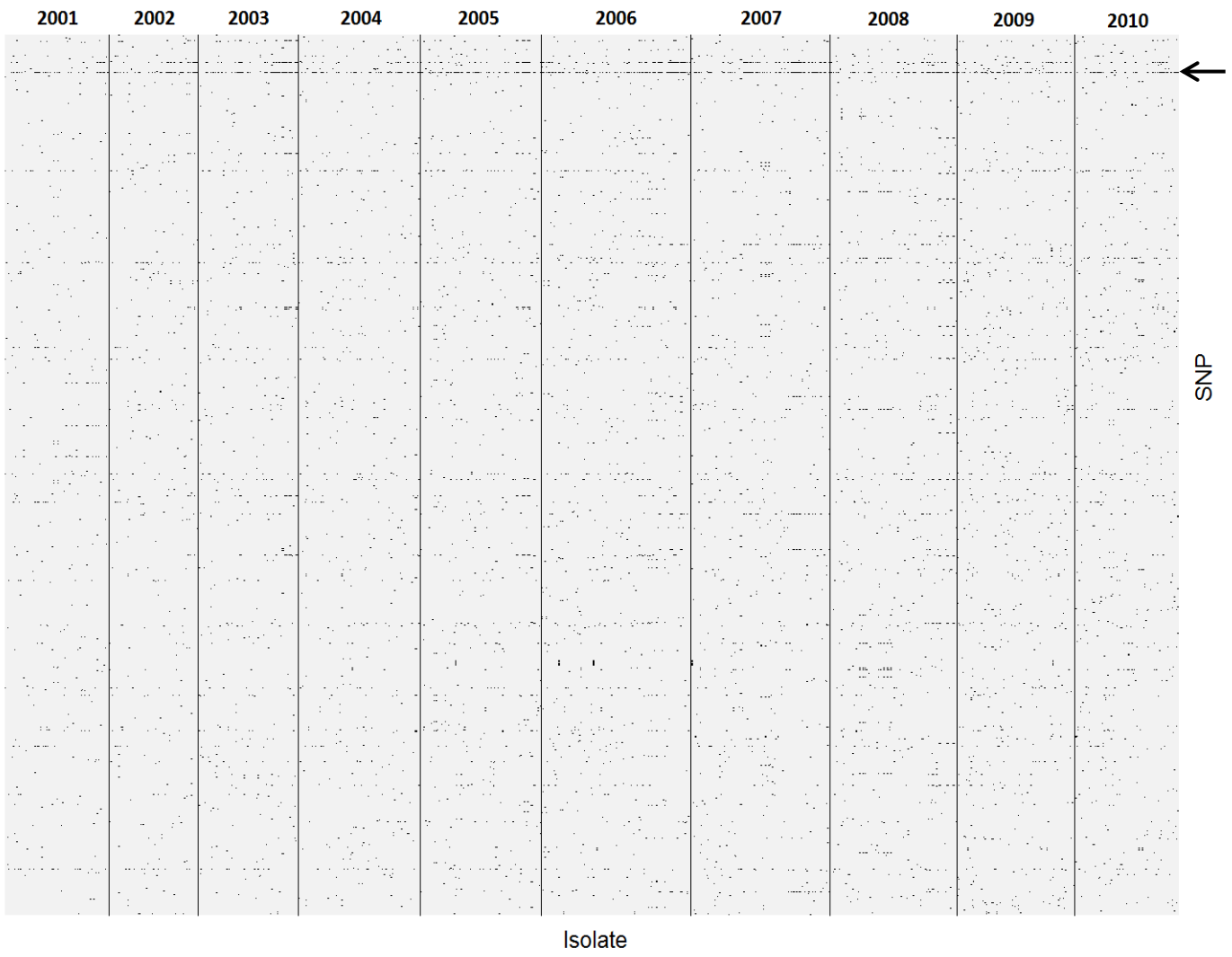


Figure 2.10. There are 5469 bi-allelic non-singleton SNP positions across the 1022 isolates. Each bi-allelic SNP position has, across the 1022 isolates, one nucleotide which is more common than the other. This can be represented as a majority nucleotide (grey) and a minority nucleotide (black). The majority nucleotide can be considered as the genetic background for that SNP position, while the minority nucleotide is the mutation that causes that position to be a SNP. This plot shows variation in the number of minority nucleotides per position. For example, there is a SNP position with a high number of minority nucleotides (black arrow). The isolates are ordered left-to-right by sampling date and the SNP positions are ordered top-to-bottom by the position on the circular genome in ascending order from the origin of replication.

There is some variation in the number of SNPs each isolate contains (Figure 2.11), with a slight trend for the later isolates to harbour more SNPs (Figure 2.12). This is as expected, since once a mutation occurs at a SNP position it is unlikely to be reversed and is often stably inherited. This stable inheritance means that the phylogenetic analysis, and the novel methods developed in future chapters, are likely to be valid approaches since a shared SNP would be indicative of shared ancestry. There is also considerable variation in the number of isolates which exhibit a particular SNP (Figure 2.13), with the majority of SNPs found in few isolates. Furthermore, there appears to be a slight trend for the more common SNPs to have been seen earlier in the Bi-allelic Dataset (Figure 2.14). This implies that there might be a steady evolutionary progression of SNPs from rare to common with eventually the minority SNP becoming the majority one. The finding that isolates sampled later have more SNPs and that the more common SNPs are seen earlier is indicative of a noticeable mutation rate. Therefore, it might be possible to track the evolution of an isolate, and determine its origin location.

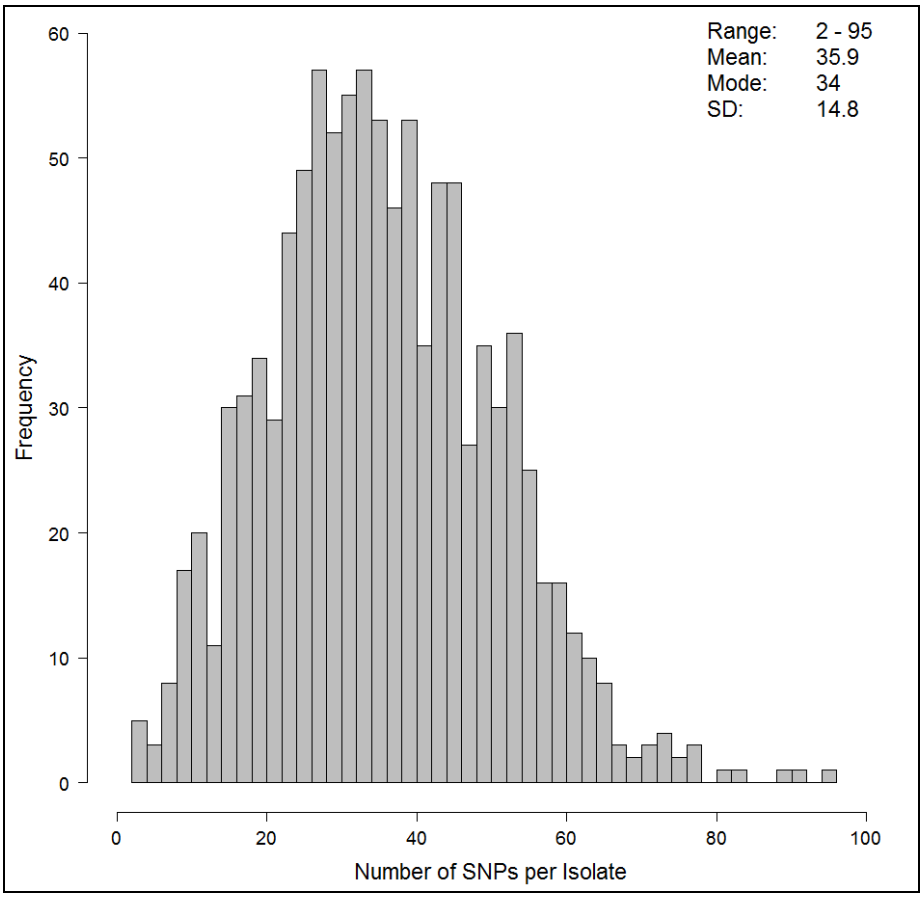


Figure 2.11. The distribution of the number of SNPs of the 1022 isolates in the Bi-allelic Dataset shows that although each isolate can vary in any of the 5469 SNP positions, the actual number of SNPs in each isolate is considerably less. Therefore an entire column of red minority nucleotides does not occur in Figure 2.10.

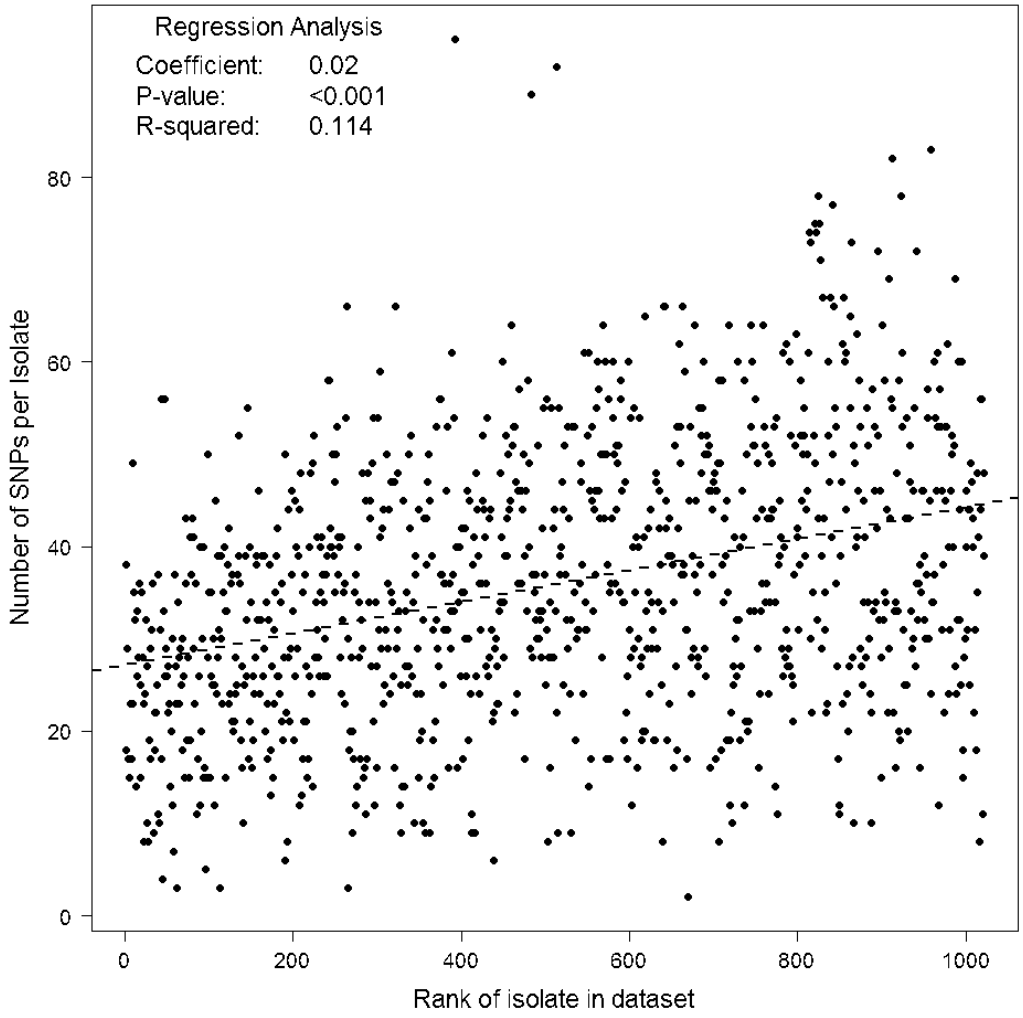


Figure 2.12. There is a slight trend (R^2 = 0.114) for isolates that are sampled later in the Bi-allelic Dataset to harbour more SNPs. This is since once a mutation occurs at a SNP position, and is not selected against, it is unlikely to be reversed and is often stably inherited. The Bi-allelic Dataset is ordered with the first isolate sampled in 2001 at Rank 1, and the last isolate sampled in 2010 at Rank 1022. Each point represents a separate isolate in the Bi-allelic Dataset.

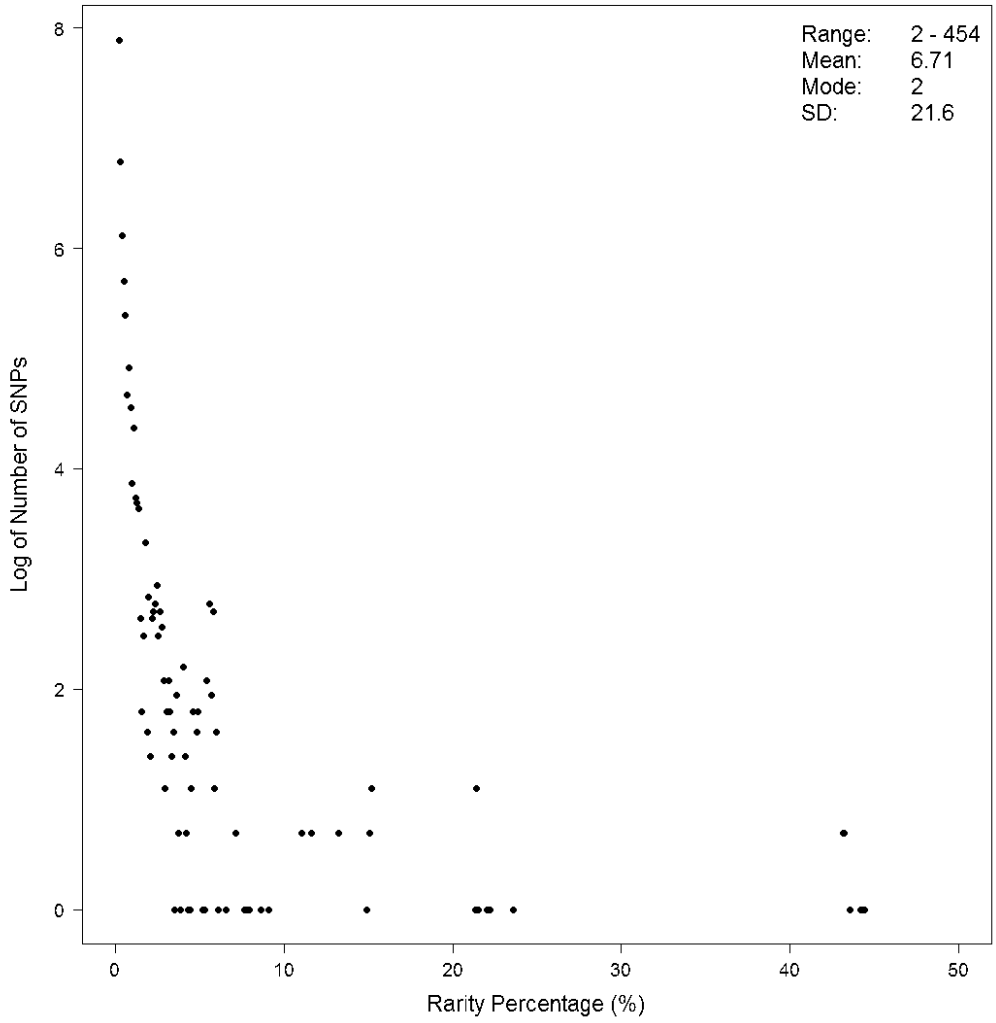


Figure 2.13. There is considerable variation in the rarity of the 5469 non-singleton bi-allelic SNPs. Rarity is defined as the percentage of the 1022 isolates which harbour that SNP. Some SNPs appear in many isolates while the majority of SNPs appear in only a few isolates. With the removal of singletons in this study a SNP can be found in a minimum of two isolates. The majority of SNPs (n = 2655) are contained in only two isolates, while the most common SNP is found in 454 isolates.

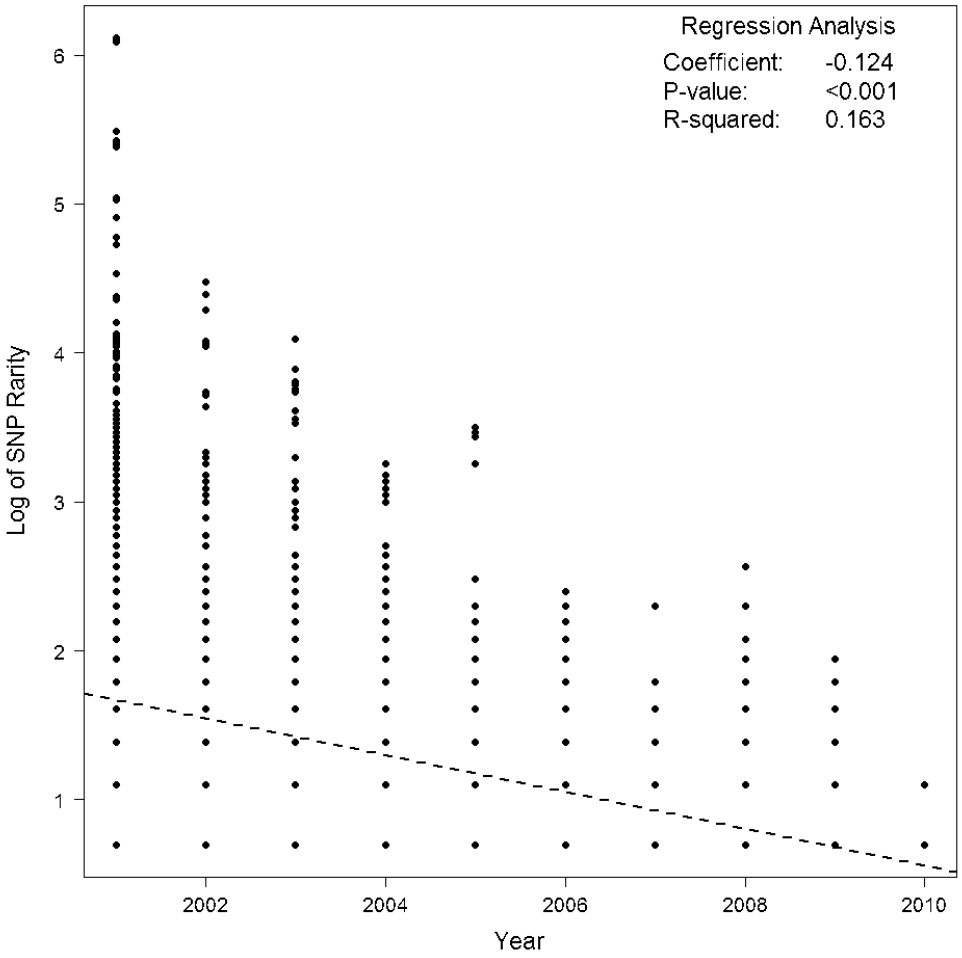


Figure 2.14. Using linear regression, there is a slight trend for the rarer SNPs to be first seen later in the Bi-allelic Dataset (Coefficient = -0.124, p-value = <0.001). Therefore, if this trend will continue there would be a progression of rare SNPs to more common ones, with eventually the minority SNP becoming the majority SNP. In this way there will be a steady turn-over of nucleotides at any given bi-allelic SNP position. Each point indicates when a SNP was first seen and in how many isolates over the 10 years in the Bi-allelic Dataset.

There is a trend for the earlier SNPs to be present in more isolates. Therefore these SNPs may have an increased geographic range. However, this posited increase in geographic range may be limited by the number of patient referrals between RCs. I hypothesised that if MRSA is spread by patient transfer then these RCs would limit the number of unique hospitals a SNP is present in, when compared to the expected number. Therefore, the number of isolates which harbour each SNP and the number of unique hospitals they were sampled from was recorded. These observed values were compared to expected values calculated by sampling at random, without replacement, successive numbers of isolates from the Bi-allelic Dataset. The number of unique hospitals that these randomly selected isolates were sampled from was recorded. This is done for 1 to 500 random isolates. This procedure was repeated

5000 times. As shown in Figure 2.15, there are a large number of SNPs which are in fewer than expected hospitals.

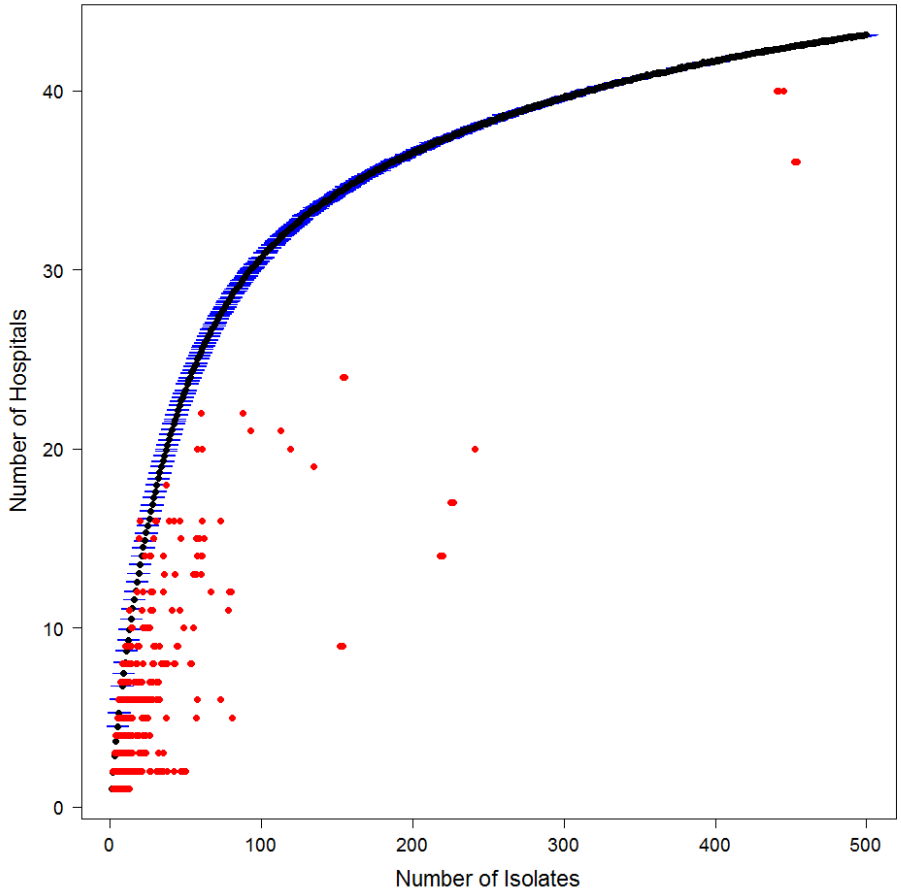


Figure 2.15. The red points indicate the number of isolates and the number of unique hospitals each individual SNP was found in. Since each SNP can only be found in one isolate a random sample of isolates from the Bi-allelic Dataset was taken (iteratively between 1 and 500 isolates) to determine the number of unique hospitals that many isolates are expected to be found in. This was repeated 5000 times. The mean number of unique hospitals (black curve) is shown with the Standard Error of the Mean (SEM; blue error bars). Many of the individual SNPs fall short of the expected number of unique hospitals they should be present in for the given number of isolates they are found in. Therefore the SNPs appear to show geographic clustering to particular hospitals, indicating that there is some process limiting the spread of the SNPs.

2.4.3 Sets of linked SNPs

Some of the 5469 SNPs are harboured in the same number of isolates and hospitals. Furthermore, some SNPs show the exact same pattern of incidence; i.e. they are always present in the exact same isolates in the Bi-allelic Dataset. Therefore, it is possible that these SNPs are linked together by some mechanism; for example, linkage disequilibrium. A group of linked SNPs is termed as a Set of Linked SNPs (SLS). Each SNP in an SLS is providing the same

repeated information. Therefore, one SNP may be used per SLS to represent the one unique piece of information obtained. This representation could be considered similar to Tag SNPs, which are used to represent a group of SNPs associated by linkage disequilibrium (Chen *et al*., 2014). It was found that there are 1391 unique SLSs in the Bi-allelic Dataset. However, there is considerable variation in the number of SNPs which form an SLS (Figure 2.16).

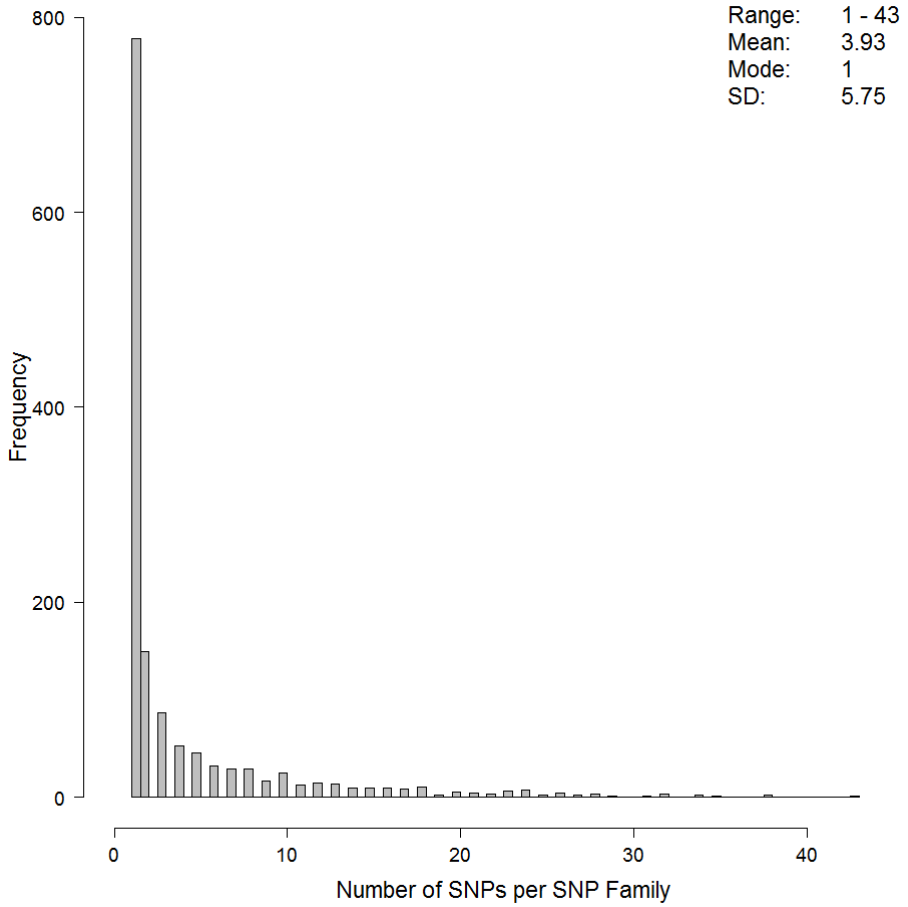


Figure 2.16. An SLS is a group of SNPs which are always harboured in the exact same isolates as another SNP in the Bi-allelic Dataset. The majority of SLSs only contain 1 SNP; i.e. there is no other SNP with the exact same incidence pattern. However, there is considerable variation in the number of SNPs in each SLS. Therefore it is possible that these SLSs contain duplicated information, which may affect the interpretation of the similarity between any two isolates.

However, the membership of each SLS can change if additional isolates are added to the dataset (either to the Bi-allelic Dataset, or prior to that in the Prime Dataset). Furthermore, although the definition of an SLS is a group of SNPS that are contained in the exact same isolates, some SLSs only differ by the absence of one SNP in one isolate. These SLSs are "nested" within each other. The SLSs can either be fully nested (e.g. SLS A is completely within

SLS B), or partially nested (e.g. SLS A is mostly within SLS B, except for a few isolates). This is visualised in Figure 2.17.

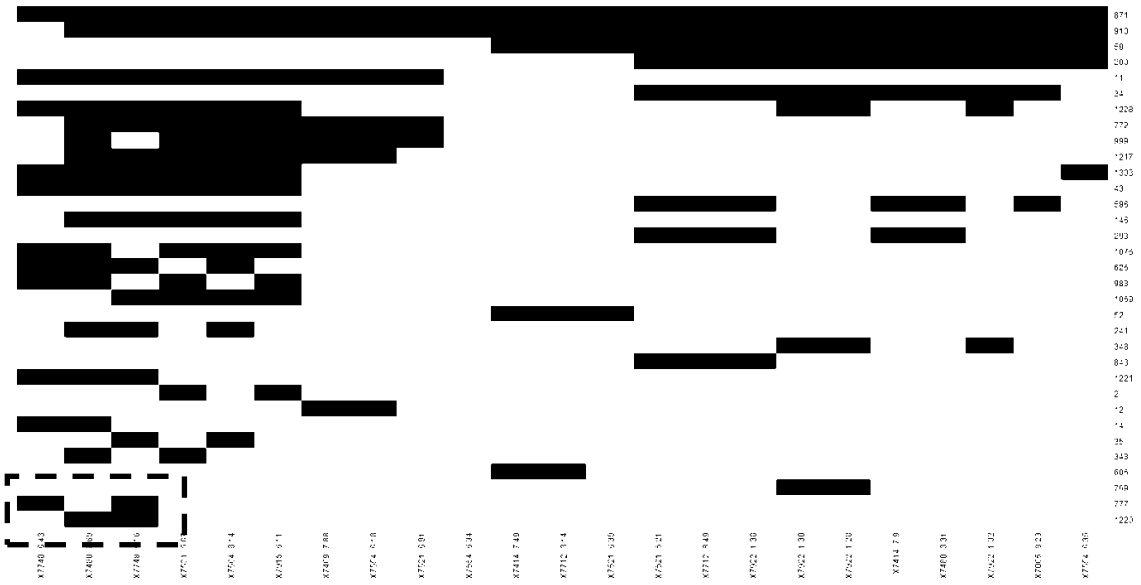


Figure 2.17. An extract from Figure 2.10 showing 23 of the 1022 isolates and 33 of the 5469 bi-allelic non-singleton SNPs. The isolates are ordered on the x-axis by sampling date and the SNPs are ordered on the y-axis according to increasing rarity. A black colour denotes a mutation in that isolate of that SNP, while the grey colour denotes the background nucleotide. Every SNP in this diagram could be considered to be nested within the top SNP. However, there are some SNPs which are only partially nested within each other, a simple example is denoted by the dashed black box.

Examining the incidence of each SLS it was found that almost every SLS is entirely or partially nested within another. This partial nestedness may be attributable to insufficient sampling, recombination of the genetic material, or separate evolution of the SNPs in different isolates since the first appearance. Therefore, using SLSs might not be a valid approach. Although some pseudo-replication may be created by using individual SNPs, each SNP is assigned equal importance. The same cannot be said for the SLSs due to their fluid definition and nestedness arising from the sampling of this collated Bi-allelic Dataset. Therefore, the conservative approach is taken and individual SNPs are used instead of SLSs in this thesis. It must be noted, that since the SLSs show nestedness, both full and partial, then the SNPs would also show nestedness. Therefore, one must remain aware of this nestedness when interpreting the phylogenetic analysis or the novel methods developed in later chapters.

2.5 Factors influencing MRSA sub-population genetic similarity

The population of MRSA examined is from the UK and Ireland. This MRSA population can be divided into two levels of sub-populations: hospital and RC. In this section I will examine the connectivity of the sub-populations, both in terms of patient referral and genetic similarity by using the Bi-allelic Dataset of 1022 isolates with 5469 bi-allelic non-singleton SNPs. The inter-connectedness of the MRSA sub-populations can be considered a network, with the nodes as individual hospitals or RCs and the edges connecting them as the geographic distance, genetic similarity or patient referral. I show that there is genetic segregation of the MRSA sub-populations based on their geographic proximity to one another. I further show that this genetic similarity may be influenced by the level of patient referral between the sub-populations and geographic proximity of the sub-populations.

In this section two statistical analyses are used: the Mantel correlation test to determine the correlation between two matrices (Mantel, 1967), and Newman's measure of assortativity (Newman, 2003) with jack-knifing (Efron & Gong, 1983). Assortativity is used to determine the preference of nodes to attach to other nodes which are similar. Similar nodes can be grouped into classes. However, in this thesis there are pre-defined classes, assigned by the geographic sampling locations (i.e. hospital or RC), where higher genetic similarity within the class than between might be expected. Therefore, the assortativity with jack-knifing will measure how well the genetic variation in the MRSA sub-populations conforms to these pre-defined classes; i.e. a measure of the level to which the sub-populations are segregating.

The direct distance and the road distance between each pair of hospitals (Appendix A Supplementary Tables A2 and A3, respectively) was found to be highly correlated ($r = 0.914$, $p = 0.001$) using a Mantel correlation test. The road distance was calculated using the shortest driving distance according to Google Maps. In this section the direct geographic distance is used as the measure of geographic proximity of MRSA sub-populations.

2.5.1 Hospital sub-population genetic similarity

The initial step was to determine if there is any effect of geographic distance on the genetic similarity of hospital sub-populations. In population genetics a conventional method to determine genetic similarity is to calculate the Fixation Index (F_{ST}; Equation 2.1; Holsinger & Weir, 2009; Meirmans & Hedrick, 2011) of each pairwise sub-population (Appendix A Supplementary Table A4). This was one of the measures used in the Ke *et al*. (2012) study which showed that patient transfers increase MRSA sub-population similarity. F_{ST} is a measure of population structure which compares the average number of pairwise genetic differences

between two sub-populations ($\pi_{BETWEEN}$) with the average number within the same sub-population (π_{WITHIN}):

$$F_{ST} = \frac{\pi_{BETWEEN} - \pi_{WITHIN}}{\pi_{BETWEEN}}. \qquad \textbf{Equation 2.1}$$

Values close to 0 indicate that the two populations are highly genetically similar, while values close to 1 imply that the two sub-populations are genetically distinct (Hudson *et al*., 1992). In certain rare cases it is possible to obtain negative F_{ST} values, which implies that the isolates from different sub-populations are genetically more similar than the isolates within a sub-population.

An increase of geographic distance would be expected to cause a decrease in the level of similarity between the MRSA sub-populations. A Mantel correlation test between F_{ST} and the direct geographic distance was conducted. It was found that there was a significant relationship ($r = 0.155$, $p = 0.047$) between the F_{ST} values and the direct geographic distance between hospitals; i.e. a higher F_{ST} value is associated with a greater distance. This indicates that geographic proximity may affect the genetic similarity of the MRSA sub-populations.

However, the use and interpretation of F_{ST} have limitations (see Meirmans & Hedrick (2011) for full review). The one most pertinent to the Bi-allelic Dataset is the issue of uneven sampling. With fewer than three samples per sub-population there is no meaningful π_{WITHIN}. In the Bi-allelic Dataset there are some locations with fewer than three samples, therefore these F_{ST} values may be incorrect. Also, F_{ST} cannot be calculated on a sub-population containing only one isolate, such as the case of Ulster Hospital in this thesis. Therefore, a different measure is needed to determine the genetic similarity of the hospital MRSA sub-populations. As previously found (Figure 2.15), certain SNPs are in fewer hospitals than expected. The variation in the incidence of SNPs could be used as the unit of similarity between sub-populations.

A new method of comparing the genetic similarity between two isolates based on their SNPs was developed. This is termed the $SNP_{SIMILARITY}$. The SNPs found in one isolate (S_i) were compared with those contained in another isolate (S_j). These isolates may be from the same location, or from different locations. Each SNP which was found in both isolates indicates increased the similarity between the two isolates. The number of SNPs found in both isolates (S_B) is standardised by the sum of the number of SNPs harboured in each isolate ($S_i + S_j$; Equation 2.2):

$$SNP_{SIMILARITY} = \frac{S_B}{S_i + S_j}. \qquad \textbf{Equation 2.2}$$

This provides a measure of $SNP_{SIMILARITY}$ for any two isolates. The values can range from 0 which indicates the two isolates have completely dissimilar SNPs, to 0.5 which indicates the isolates have identical SNPs. One isolate was randomly selected from each hospital in a pair and calculated the $SNP_{SIMILARITY}$ measure. This was repeated 5000 times for each pair of hospitals. The mean of the 5000 repeats is the value of the SNP similarity between the two hospitals. This was repeated for all 46 hospitals in the Bi-allelic Dataset to populate a $SNP_{SIMILARITY}$ matrix (Appendix A Supplementary Table A5). A higher SNP similarity value would indicate a closer genetic association between the sub-populations. Therefore, it might be expected that with increasing geographic distance there will be a corresponding decrease in SNP similarity. It was found that there is a significant negative relationship between these two measures (Mantel correlation $r = -0.1571$, $p = 0.003$). This finding further supports the geographic segregation of the sub-populations.

Another alternative measure of genetic similarity between MRSA sub-populations was developed for this project. This is termed the $SNP_{CONNECTIVITY}$. Each SNP was examined in turn and identified the sub-populations which contained the SNP. These sub-populations are therefore connected via the sharing of this SNP and so this increases their genetic similarity measure. For example, if a SNP is seen in two hospitals (A and B) then this is considered a connection between these two hospitals. Furthermore, if multiple isolates in the same sub-population harbour the SNP this will count towards their self-similarity. For each pair of hospitals it was determined which of the 5469 SNPs appear in both. This number is now the $SNP_{CONNECTIVITY}$ of those two hospitals. This procedure was repeated for all 46 hospitals and populated a $SNP_{CONNECTIVITY}$ matrix (Appendix A Supplementary Table A6). A higher number of SNPs appearing in both hospitals indicates a greater level of connectivity between the sub-populations. The SNPs found in multiple isolates in the same hospital are used to give a measure of the assortativity of the matrix.

Using the $SNP_{CONNECTIVITY}$ matrix it was found that there is a hospital-level assortativity of $r = 0.159$ ($p < 0.001$). However, the majority of SNPs are only ever seen in two isolates (Figure 2.13). These SNPs can be termed “rare SNPs”. These rare SNPs might be more informative since the more common SNPs might be present in all sub-populations, and therefore contribute “noise” to the network. Therefore any SNP was removed which was found in more than two isolates, and created a Rare SNP Dataset which comprises 1022 isolates and 2655 SNPs. Using the Rare SNP Dataset there is a much greater level of assortativity (Newman’s $r = 0.609$, $p < 0.001$; Appendix A Supplementary Table A7). This is supported by a Mantel correlation test which shows that there is a significant negative

relationship between the direct geographical distances of the hospitals and the level of SNP connectivity, using either the Bi-allelic Dataset (r = -0.2183, p = 0.001) or just the Rare SNP Dataset (r = -0.1499, p = 0.001). This implies that with increased geographic distance between two hospitals there is reduced genetic similarity in the sub-populations.

All three genetic similarity measures (F_{ST}, $SNP_{SIMILARITY}$, and $SNP_{CONNECTIVITY}$) support the conclusion that the geographic proximity of the MRSA sub-populations influences the genetic similarity. However, this relationship is relatively weak with r-values around 0.15. Therefore, the geographic proximity of the sub-populations is not the only factor influencing genetic similarity.

2.5.2 Patient referral influences hospital sub-population genetic similarity

It is possible that it is not solely the geographic distance between hospitals which is important, but rather some other measure of connectivity. A UK study by Donker *et al.* (2012, 2014) posited that it is the level of patient referral between hospitals which influences the connectivity. This was also the finding by Ke *et al* (2012) in hospitals from California and by Donker *et al*. (2010) in Dutch hospitals. If this is the case then it is possible for a pair of hospitals which are geographically distant to become genetically closer with increased number of patient referrals. Although Donker *et al.* (2012, 2014) contained samples from more English hospitals than are present in this thesis, I investigated whether the SNP connectivity network created here would still support the RC definitions. For the comparison of patient referral data and the genetic similarity the $SNP_{CONNECTIVITY}$ measure described in the previous section was used.

Since Donker *et al*. (2012, 2014) only used patient referral data from English hospitals all sampling locations were excluded where there was not also patient referral data. This leaves 27 hospitals from England where there is both patient referral data and genetic data. The relevant patient referral data was extracted for the hospitals in this study (Appendix A Supplementary Table A8) and it was found that there remains a high level of patient referral assortativity at both the hospital level (Newman's r = 0.976, p < 0.001) and the RC level (Newman's r = 0.993, p < 0.001) for this England-only set of locations. Therefore, even on this attenuated sub-set of the Donker *et al.* (2012, 2014) study, the patient referral data still conforms to the RC definitions.

I next confirmed, on this England-only subset of the Bi-allelic Dataset, that there is still assortment of the SNPs by RCs. When including all the bi-allelic SNPs there is a lower level of

assortment of SNPs by RC (Newman's r = 0.304, p = 0.003) than when only using those rare SNPs (Newman's r = 0.686, p < 0.001). This implies that there is an effect of the RCs on the SNP similarity of RC sub-populations. Therefore, even with the much fewer English hospitals available in this thesis the RCs are still a valid delineation of geographic regions.

Therefore, to determine the influence of patient referral on genetic similarity of MRSA sub-populations, a Mantel correlation analysis was conducted. A significant correlation was discovered between the patient referral data and the $SNP_{CONNECTIVITY}$ data at the hospital level, using either all the bi-allelic SNPs (r = 0.158, p = 0.01) or just the rare SNPs (r = 0.220, p = 0.013). At the RC level there is an even stronger correlation between the patient referrals and the $SNP_{CONNECTIVITY}$, using either all the bi-allelic SNPs (r = 0.558, p = 0.001) or just the rare SNPs (r = 0.310, p = 0.041). This indicates that a higher number of patient referrals is related to an increase of genetic similarity between the MRSA sub-populations at both the hospital level and the RC level. There are higher r-values in the Mantel correlation tests between the patient referral data and the genetic similarity than between the geographic distance and genetic similarity. Therefore, it appears that patient referrals are a more important factor in determining genetic similarity of MRSA sub-populations. It would be important for all hospitals within a referral cluster to adopt the same infection prevention strategies, and those which might be considered hubs (e.g. large city hospitals) to have targeted prevention measures (Ciccolini *et al*., 2013). Additionally, as was concluded by Donker *et al.* (2010), the rate of patient transfers should be included when using the rate of MRSA incidence to gauge the level of hospital hygiene.

These findings show that the MRSA sub-populations exhibit genetic diversity based on their geographic proximity. Further than that, there is population structure based around the amount of patient referrals between hospitals and between the RCs posited in Donker *et al.* (2012, 2014). Analysis of the genetic similarity of the MRSA sub-populations indicate that these RCs are a valid way to geographically segregate the hospitals in this thesis. Therefore, there are two geographic resolutions for attempting to determine the geographic origin of an isolate in future chapters; the hospital resolution, and the coarser RC resolution. Furthermore, the genetic differentiation of MRSA sub-populations might allow for the indentification of transmission events.

2.6 Conclusion

Although there is variable sampling in the data used in this thesis, phylogenetic examination of the Prime Dataset indicates that there may be geographic clustering of genetically similar isolates. Using the Bi-allelic Dataset I show that there is variation in the rarity of the SNPs, with some SNPs found in fewer locations than expected for their relative abundance. Although geographic proximity plays a role in limiting SNPs to specific locations, I show that the level of patient referrals between locations is an important factor affecting the genetic similarity of the sub-populations.

Therefore I have shown that there is some evidence for the movement of genetic material from one MRSA sub-population to another. The next step would be to determine which of the isolates between any two pairwise sub-populations are causing the increased genetic similarity. This identification of isolates that have been introduced to a new sub-population (i.e. transmission events) could provide information on the spread of particular strains or virulence types of MRSA.

3 Identification of MRSA introduction events

3.1 Background

Phylogenetic analysis is an integral part of the process to determine the geographic origin of a pathogen, which helps to understand the epidemiology, transmission rates, and the most effective control measures. Phylogenetic analysis is often used to determine the transmission route and origin of an MRSA isolate or lineage (for example, Deurenberg *et al*., 2005; Harris *et al*., 2010; Harrison *et al*., 2013; Holden *et al*., 2013; McAdam *et al*., 2012). However, there are limitations to the practicality of using a phylogenetic approach. For example, increasing the number of isolates or amount of genetic information per isolate drastically increases the computational time required (Kuhner & Felsenstein, 1994). Therefore, I investigated whether there may be an alternative route to determine the geographic origin of a pathogen.

In the previous chapter I demonstrated a genetic differentiation between the MRSA sub-populations based on the Single Nucleotide Polymorphisms (SNPs), both at the Referral Cluster (RC) or hospital geographic level of resolution. This structure in the genetic similarity may be attributable to geographic proximity and level of patient referrals between sub-populations. Furthermore, the SNPs are not uniformly distributed across all isolates (Section 2.4.2) and it is possible that certain SNPs may be contained within a particular geographical area. Therefore, I investigated if any of the SNPs provided a signature for a particular geographic location and if this information can be used to identify transmission events of MRSA from one sub-population to another. These will be termed Candidate Introduction (CI) events. Any SNP shared between isolates is assumed to be an indication of a shared ancestor of those isolates. This assumption can be made due to the clonal nature of MRSA that results in a very low rate of homoplasy and recombination in the core genome (Castillo-Ramírez *et al*., 2011). Furthermore, the SNPs known to be associated with drug-resistance genes were removed to reduce the incidence of homoplasy (see Section 2.4.1).

A signature SNP for a geographic location would be one which is only ever found in isolates sampled in that geographic location. For example, if a SNP is only found in isolates

from Cambridge, then this SNP is a signature for Cambridge. If any subsequently sampled isolates harbour this SNP it could be an indication of an association with Cambridge. Therefore, this signature SNP can be used as a diagnostic to indicate the potential origin of that isolate as Cambridge. Although I focus on the geographic origin characteristic in this chapter, it is important to note that signature SNPs might be present for other characteristics provided the appropriate metadata are known. For example, if a certain SNP is always harboured in isolates with a certain virulence factor then this SNP would be a signature SNP for that virulence factor.

In Section 2.4.2 it was found that many of the SNPs in the Prime Dataset are singletons; i.e. only ever found in one isolate. Singletons, although useful in phylogenetic analysis to determine branch length, will not be signature SNPs since they do not provide information on the similarity between isolates. However, singletons do provide a large source of potential signature SNPs if more isolates were to be added to the Prime Dataset. If a singleton SNP is subsequently seen in a newly added isolate which shares the same characteristic (e.g. geographic location), this SNP is now a signature SNP. In this manner there would be a continuous turn-over of SNPs; from singletons, to possible signature SNPs, to becoming more common (see Chapter 2, Figure 2.11). Therefore, although in this chapter the Bi-allelic Dataset is used, if more isolates were to become available then all SNPs would need to be included. The identification of the SNPs used in this thesis is described in Section 2.2.

In summary, in this chapter I investigated introduction events based on a signature SNP. I used the Maximum Likelihood (ML) phylogenetic tree to identify possible CIs. I then determined which isolates exhibit signature SNPs. I used this information to identify CIs which may be characterised by a signature SNP. I discovered that it is possible to identify a number of isolates as CIs based on a signature SNP.

3.2 Method

3.2.1 Phylogenetic candidate introductions identification

The ML tree was used to determine any potential CIs. The first step in determining CIs was to manually partition the phylogenetic tree into sub-clades, within which the isolates could be considered phylogenetic neighbours. This process is described in detail in Section 2.4.1.

As seen in Section 2.4.1 (Figure 2.5) there are many different variations in the geographic grouping of isolates within sub-clades; from isolates all sampled in one geographic location to all isolates from disparate geographic locations (Figure 3.1). A particular type of sub-clade clustering could be used to determine if an isolate is a possible introduction event from one geographic location to another. A single isolate phylogenetically clustered with isolates from a single different geographic sampling location is suggestive of an introduction event (Figure 3.1b). The geographic location of the majority of the isolates in the sub-clade is posited to be the origin location. Since the phylogenetic tree is being used to determine the origin of the isolate this is termed as a Tree-based Assignment of Pathogen Origin (TAPO).

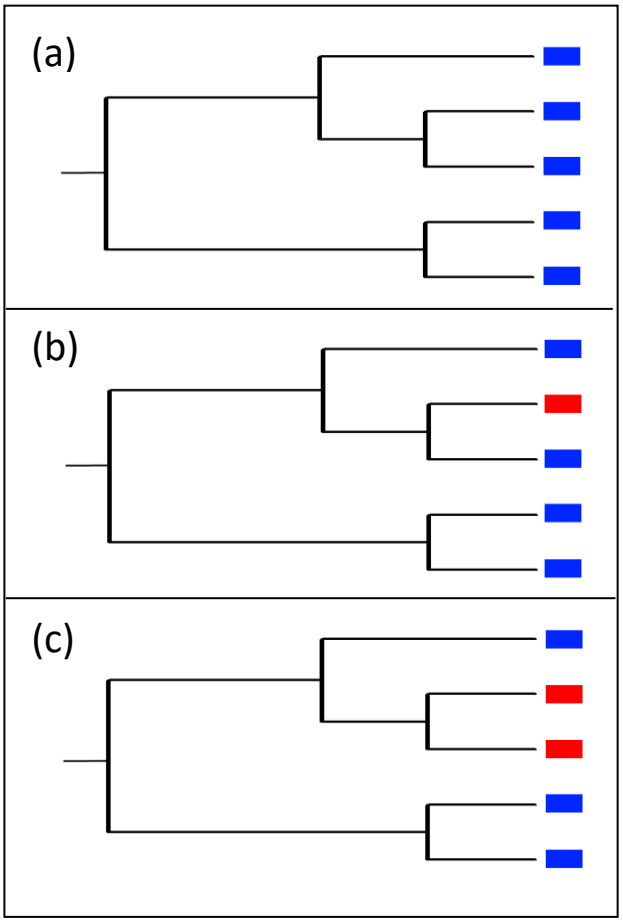


Figure 3.1. Example sub-clades of a phylogenetic tree which show phylogenetically close isolates (i.e. isolates with high genetic similarity to each other). In (a) all isolates were sampled in the blue geographic location. Therefore, it can be posited that all isolates have originated from this geographic location. In (b) all bar one of the isolates were sampled in the blue geographic location. Therefore, the isolate sampled in the red geographic location is a possible introduction from the blue geographic location. In (c) there are two possible isolates which could be introduction event. This could be indicative of either a single introduction which replicated or two separate introductions. In situations like this the conservative approach is to consider this as only one introduction event.

There are differing degrees of confidence in a phylogenetic approach when determining the origin of an isolate and whether or not it could be an introduction event. For example, if there is a large sub-clade with all isolates bar one originating from the same geographic location then this isolate is more likely to be an introduction event than the same situation in a smaller sub-clade. Alternatively, there may be a sub-clade where it is unclear which is the origin location since there are many different geographic locations represented. Therefore, stringent thresholds were applied that need to be satisfied, as follows.

Firstly, in some situations there are duplicated CIs in the same sub-clade (Figure 3.1c). In these situations it is difficult to determine if this is indicative of multiple separate introductions or one introduction that has since replicated within the new location. The conservative approach is to reduce these situations to the earliest sampled isolate and use this isolate as the possible CI. Secondly, those CIs where there is no clear geographic origin in the sub-clade due to many different locations being represented were removed. Finally, those CIs that come from sub-clades which are too small were eliminated. For this study, I only considered those CIs which are non-duplicates and come from sub-clades that have 80% or more of the isolates from one location, with a minimum of 5 isolates in a sub-clade.

There is now a method by which it may be possible to identify CIs using a phylogenetic tree. The next step is to investigate if some of these CIs can also be identified using signature SNPs. To do this, it must be determined which isolates harbour signature SNPs for a specific location.

3.2.2 Determining signature SNPs

The SNPs in the Bi-allelic Dataset are not uniformly distributed across all isolates (see Section 2.4.2), and it is possible that some SNPs will be found only in isolates from a specific geographic sampling location. Geographic locations are examined at two resolutions; individual hospitals and Referral Clusters (RCs).

The first step is to identify a focal isolate to be examined. Any isolate sampled after the focal isolate is excluded from the analysis. The focal isolate's SNPs were then identified. Each SNP was examined and it was determined in which previous isolates it was found, up to but excluding the focal isolate. If the SNP is present in isolates which have only been sampled in one location then this SNP is a signature SNP for that location. These SNPs are termed as Location Specific SNPs (LSSs). For the focal isolate any LSSs were noted. It is possible for a single isolate to have multiple LSSs for the same or different locations. This process was

repeated for all 1022 isolates in the Bi-allelic Dataset, at both the hospital and RC geographic resolution.

3.2.3 Candidate introductions with location specific SNPs

It is now known which isolates could be considered CIs and which isolates harbour LSSs. Therefore, those isolates which fall into both categories are extracted. However, it is possible that a CI may have LSSs for multiple geographic locations that contradict one another, or an LSS for a location that contradicts the TAPO posited origin. Since this is a proof-of-principle investigation to determine if an isolate can be defined as a CI based on a single SNP, those CIs which have LSSs for multiple geographic locations and those with LSSs for geographic locations than contradict the origin posited by TAPO were removed. This left those CIs that only harbour one LSS, or multiple LSSs for the same location. Furthermore, this LSS location must be the one posited as the possible origin location of the isolate by TAPO. This idea is developed in greater detail in Chapter 4.

3.3 Results

3.3.1 Phylogenetic candidate introductions determined by TAPO

The Candidate Introductions (CIs) were determined based on the ML phylogenetic tree. There were 127 possible CIs at the hospital geographic resolution, while there were 72 CIs at the RC resolution. Subsequently, those CIs which do not conform to the stringent thresholds described in the method Section 3.2.2 were excluded. Those isolates which were non-independent introduction events (see Figure 3.1c) were eliminated. Further, those isolates which came from sub-clades that were too small (i.e. less than 5 isolates) or did not show a single majority origin location were eliminated. This process resulted in 57 hospital CIs (Figure 3.2) and 39 RC CIs (Figure 3.3) which are independent and arise in sub-clades of sufficient size which have a single majority origin location.

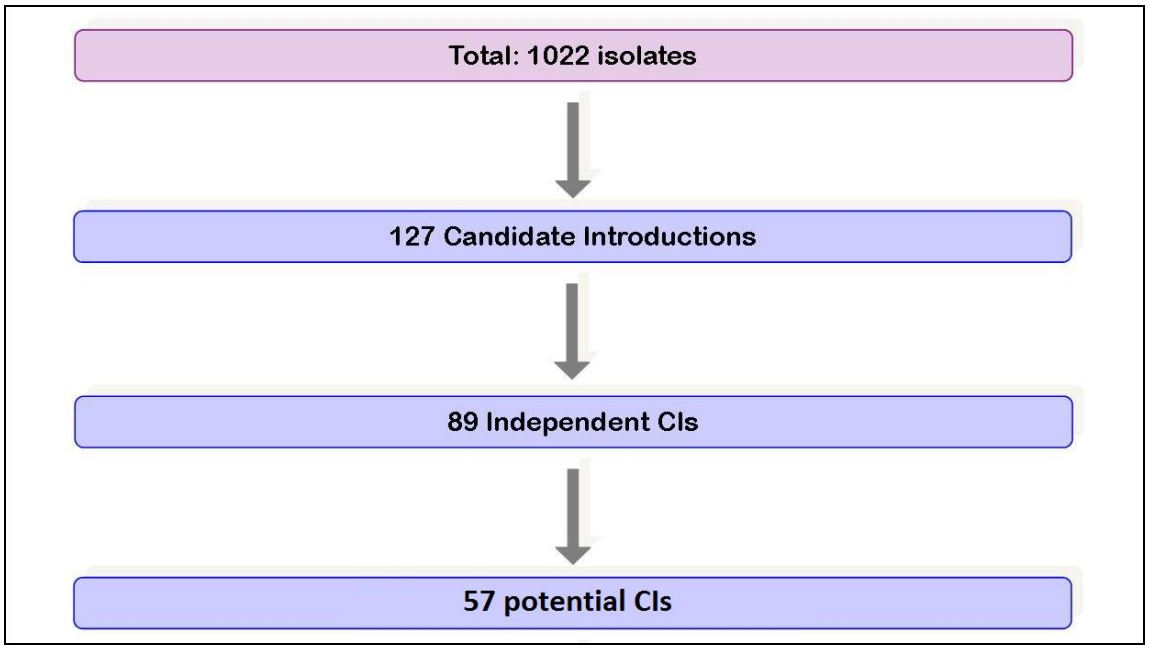


Figure 3.2. The ML phylogenetic tree was used to identify 127 possible CIs at the hospital resolution out of the 1022 isolates in the Bi-allelic Dataset. Those CIs which are not independent CIs and those without sufficient confidence were removed. This resulted in 57 potential hospital CIs based on the ML phylogenetic tree.

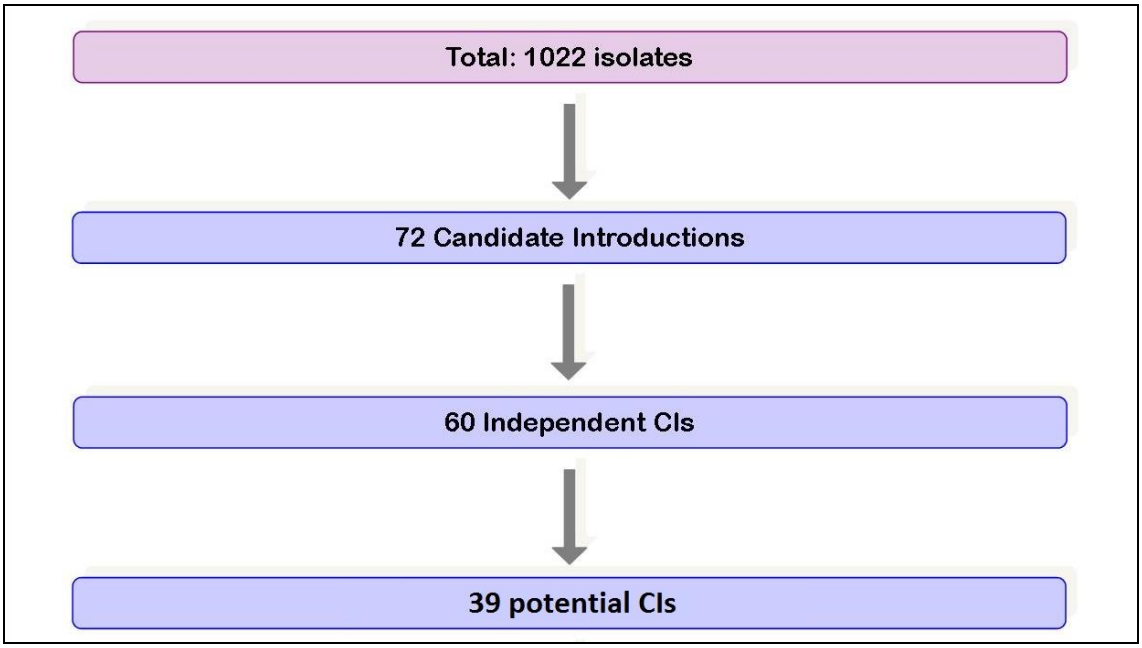


Figure 3.3. The ML phylogenetic tree was used to identify 72 possible CIs at the RC resolution out of the 1022 isolates in the Bi-allelic Dataset. Similarly to the hospital resolution CIs, those CIs which are not independent CIs and those without sufficient confidence were removed. This resulted in 39 potential RC CIs based on the ML phylogenetic tree. There are fewer CIs than at the hospital resolution due to the coarser geographic scale involved.

3.3.2 Isolates with location specific SNPs

The next step in this investigation was to identify any isolates which harbour Location Specific SNPs (LSSs). Of the 1022 isolates it was found that 881 isolates (86.2%) in the Bi-allelic Dataset harbour at least one LSS at the hospital geographic resolution (Figure 3.4).

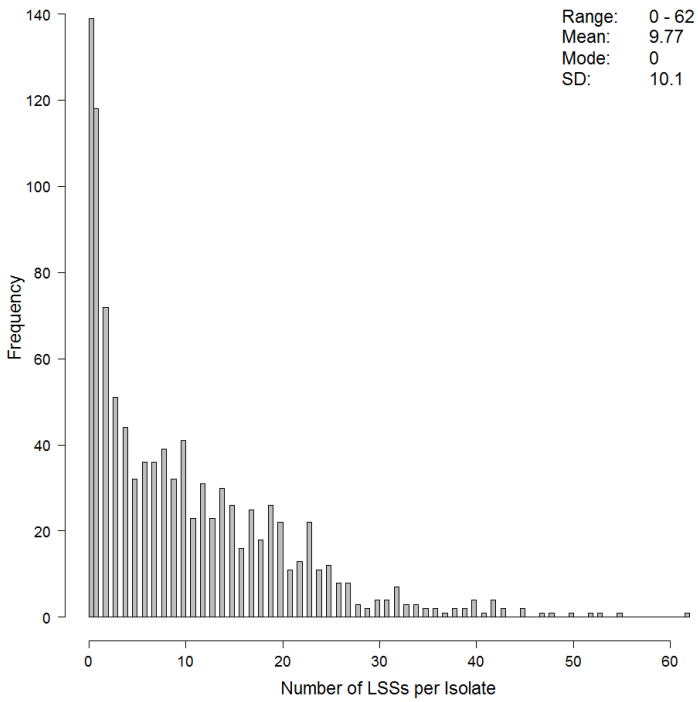


Figure 3.4. There is considerable variation in the number of hospital LSSs harboured by a given isolate. The majority (n = 881, 86.2%) of isolates have at least 1 hospital LSS, with a large number (n = 764, 74.8%) of isolates containing multiple hospital LSSs. Some isolates harbour multiple LSSs for the same hospital location. There does not appear to be any correlation between the sampling date and number of LSSs an isolate contains.

Of the 881 isolates which harbour hospital LSSs, the majority (n = 599, 68%) exhibit only one LSS location, while the rest (n = 282, 32%) exhibit multiple LSS locations. Although a slight majority of the 881 isolates harbour an LSS for the location they were sampled in (n = 490, 55.6%), an even greater number of isolates harbour an LSS for the sampling location of one of their phylogenetic neighbours (n = 652, 74.0%). Finally, a small number (n = 102, 11.6%) of the isolates which exhibit hospital LSSs were not assigned to any sub-clade on the ML phylogenetic tree (Figure 3.5).

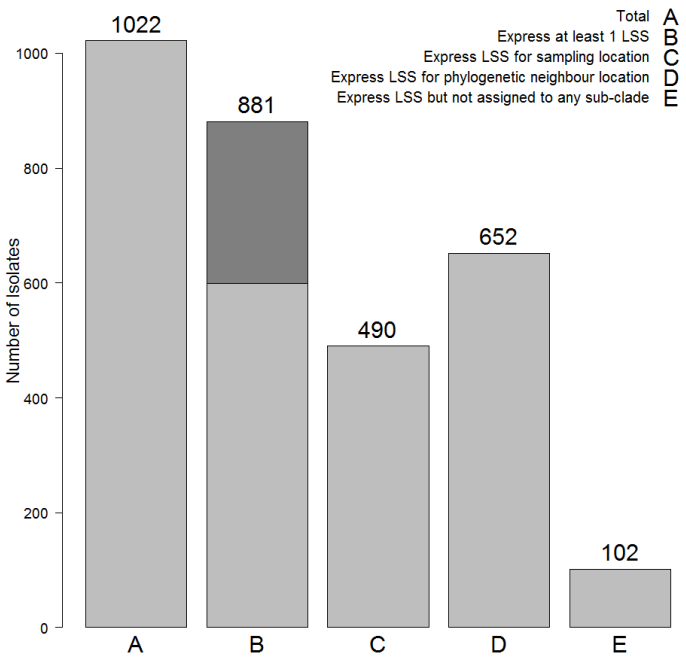


Figure 3.5. The breakdown of the hospital LSSs for the 1022 isolates. The majority of isolates contain LSSs (column B, n = 881, 86.2%), with most isolates exhibiting only one LSS (column B light grey, n = 599, 68.0%). Many of the isolates have an LSS for their sampling location (column C, n = 490, 55.6%), but a greater number have an LSS for the sampling location of at least one of their phylogenetic neighbours (column D, n = 652, 74.0%). Furthermore, a few isolates could not be assigned to any sub-clade on the ML phylogenetic tree yet still harbour LSSs (column E, n = 102, 11.6%).

A coarser resolution of geographic delineation was examined: the Referral Clusters (RCs). It was found that 924 of the 1022 isolates (90.4%) in the Bi-allelic Dataset harbour an LSS at the RC geographic resolution (Figure 3.6).

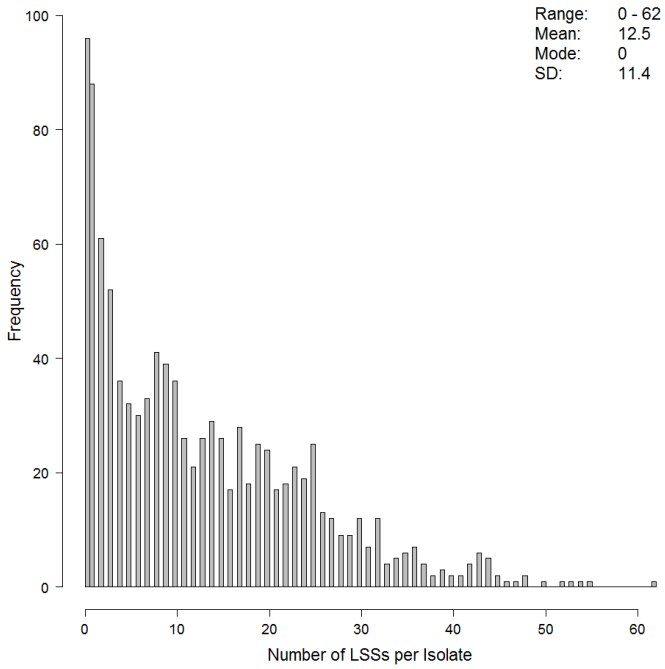


Figure 3.6. Similarly to what was found using the hospital geographic resolution, there is variation in the number of LSSs harboured by a given isolate at the RC geographic resolution. The considerable majority of isolates (n = 924, 90.4%) exhibit at least one RC LSS, with a large number of isolates (n = 836, 81.8%) harbouring multiple LSSs. Some isolates exhibit multiple LSSs for the same RC location. There does not appear to be any correlation between the sampling date and number of LSSs an isolate contains, even at this coarser level of geographic resolution.

Of the 924 isolates which do harbour RC LSSs, the majority (n = 618, 66.9%) exhibit only one LSS location, while some exhibit multiple LSS locations (n = 306, 33.1%). The majority of the 924 isolates contain an LSS for the location they were sampled in (n = 651, 70.5%). An even greater number of the isolates have an LSS for the sampling location of one of their phylogenetic neighbours (n = 728, 78.8%). Finally, a small number (n = 109, 11.8%) of the isolates which have RC LSSs were not assigned to any sub-clade on the ML phylogenetic tree (Figure 3.7).

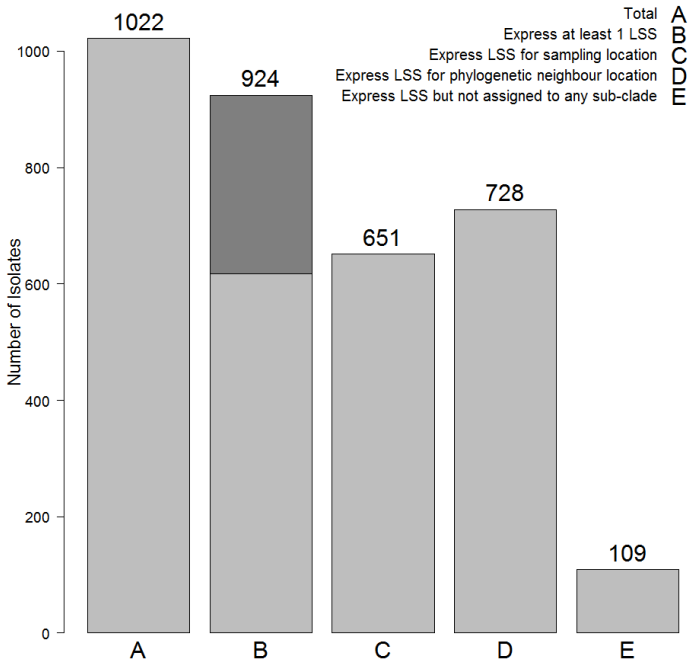


Figure 3.7. The breakdown of the RC LSSs for the 1022 isolates. As was found at the hospital geographic resolution, the majority of isolates contain LSSs (column B, n = 924, 90.4%), with most isolates harbouring only one LSS (column B light grey, 66.9%). Many of the isolates have an LSS for their sampling location (column C, 70.4%), but a greater number exhibit an LSS for the sampling location of at least one of their phylogenetic neighbours (column D, 78.8%). Furthermore, a few isolates could not be assigned to any sub-clade on the ML phylogenetic tree yet still exhibit LSSs (column E, 11.8%). The pattern of LSS incidence appears to be similar at both geographic resolutions.

3.3.3 Candidate introductions with location specific SNPs

Thus far, 57 isolates at the hospital geographic resolution and 39 isolates at the RC resolution were identified which could be considered CIs. However, these isolates show a wide variety of different LSS incidences. Since the aim is to determine if an introduction event may be characterised by a single SNP all those CIs which do not have an LSS, and those which harbour an LSS for a location other than the one posited by TAPO were excluded. Finally, those CIs which exhibit multiple LSS locations were excluded. This leaves CIs which harbour LSSs for only one location, which is the same location as that posited to be the origin by TAPO. There are 18 hospital CIs (Figure 3.8) and 16 RC CIs (Figure 3.9) that also can be characterised by an LSS for the posited origin location.

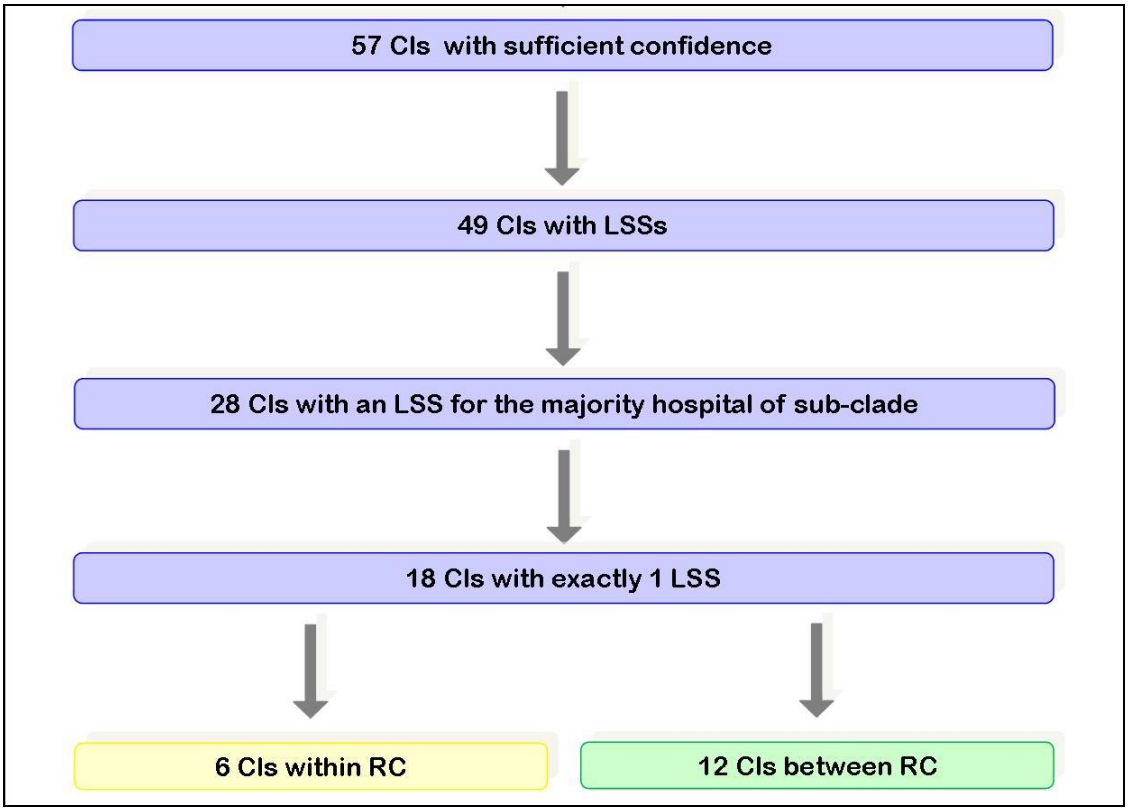


Figure 3.8. 57 possible hospital CIs were previously identified using the TAPO process. This number was further reduced based on the LSSs expressed in each CI. Those few CIs which do not contain any LSSs (n = 8) were removed. Next, those CIs which do exhibit an LSS but do so for a different location than that posited as the origin location by TAPO (n = 21) were removed. Finally, those CIs which exhibit multiple LSS locations (n = 10) were removed. This process results in 18 hospital CIs. There are a greater number of hospital CIs between RCs than there are within RCs.

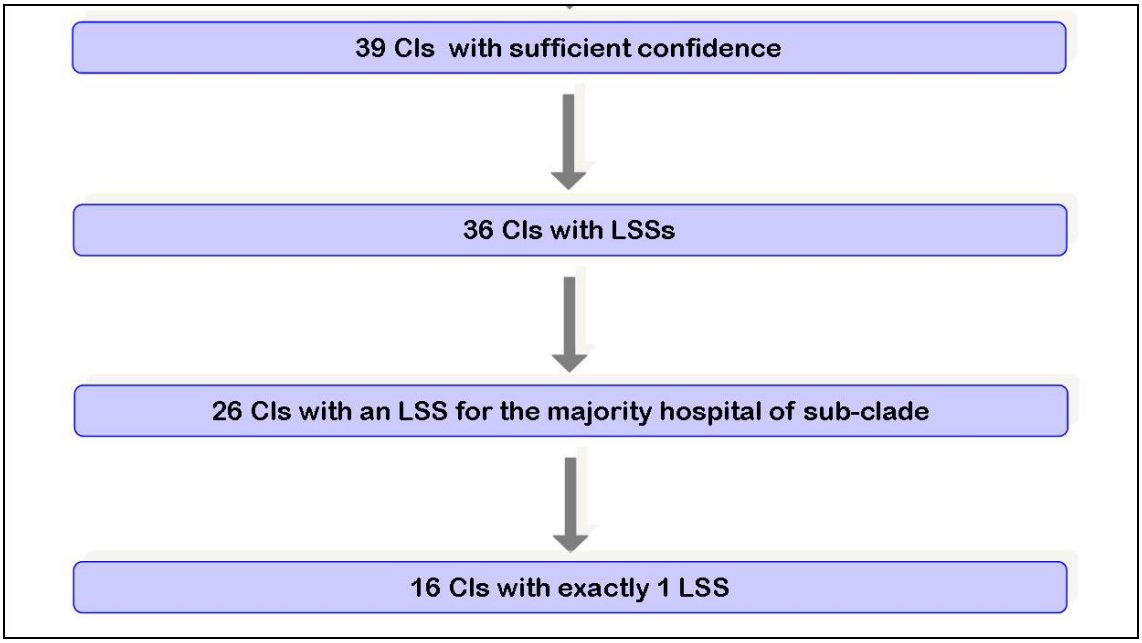


Figure 3.9. 39 possible RC CIs were previously identified using the TAPO process. The possible CIs were further reduced based on their LSSs. 3 CIs were removed which do not have any LSSs. Then, those which harbour an LSS for a different location than that posited as the origin location by TAPO (n = 10) were removed. Finally, those CIs which exhibit multiple LSS locations (n = 10) were removed. This process results in 16 RC CIs. The slight discrepancy between the hospital CIs and the RC CIs is due to the sub-clade clustering and initial definition of a CI.

The majority of the hospital CIs are between RCs (n = 12, 66.7%, Figure 3.10, Appendix B Supplementary Table B1). Examining the RC CIs, if those Cis which are between RCs that do not share a land border are ignored then the majority of RC CIs are to non-adjacent RCs (n = 10, 62.5%, Figure 3.11, Appendix B Supplementary Table B2).

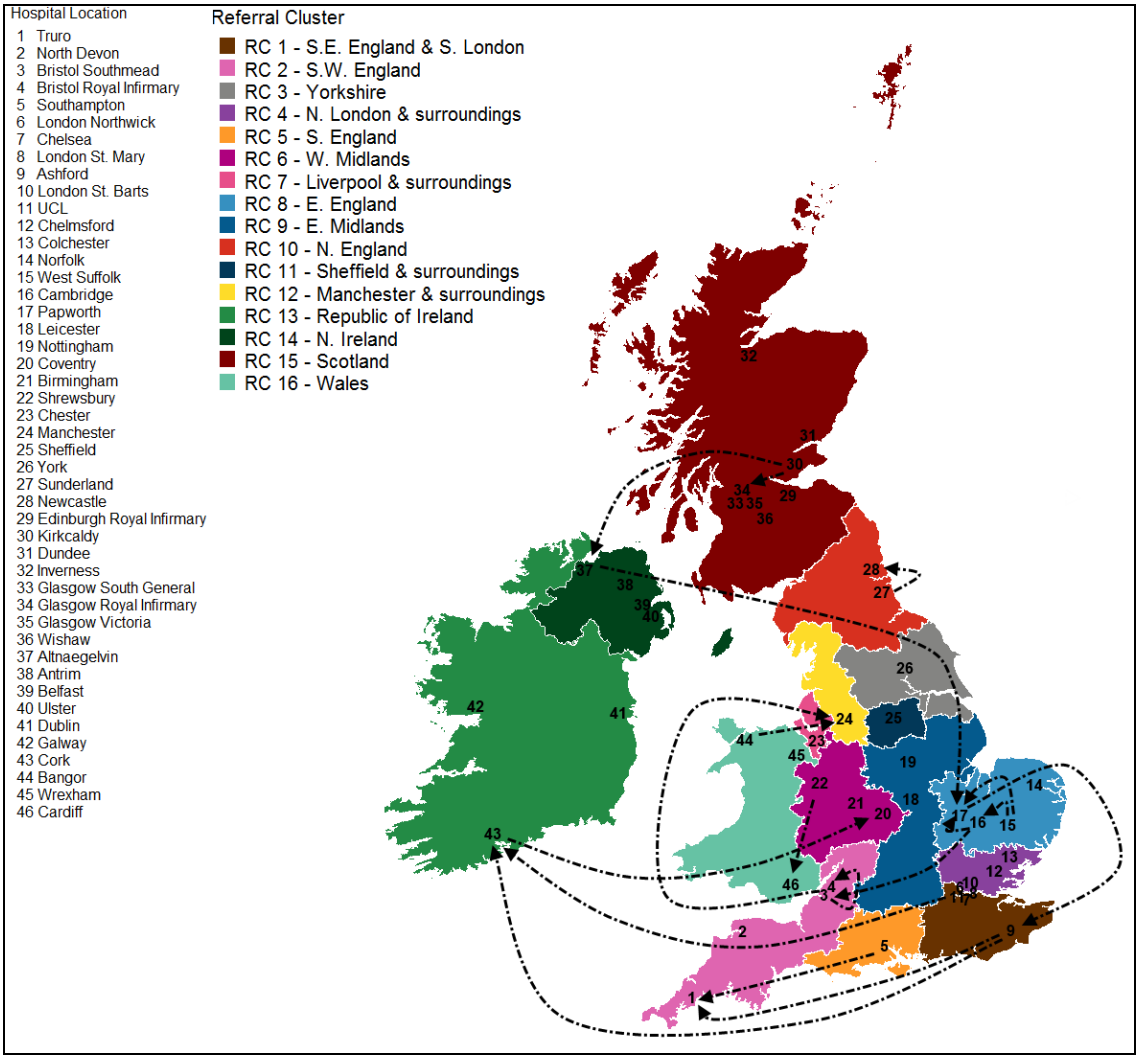


Figure 3.10. Out of 1022 isolates there are 18 hospital CIs that meet all the conservative thresholds. 127 isolates were initially identified using the TAPO process. These were then reduced to 18 CIs which have only one LSS for the same location that the TAPO method posits as the origin. There are a greater number of CIs between than within Referral Clusters (RCs), and a large number of those are to non-adjacent RCs. The hospitals have been grouped into 16 separate RCs based on the regions identified in Donker *et al.* (2012, 2014). The hospitals are numbered according to geographic location and RC, with hospitals located in the same RC grouped together.

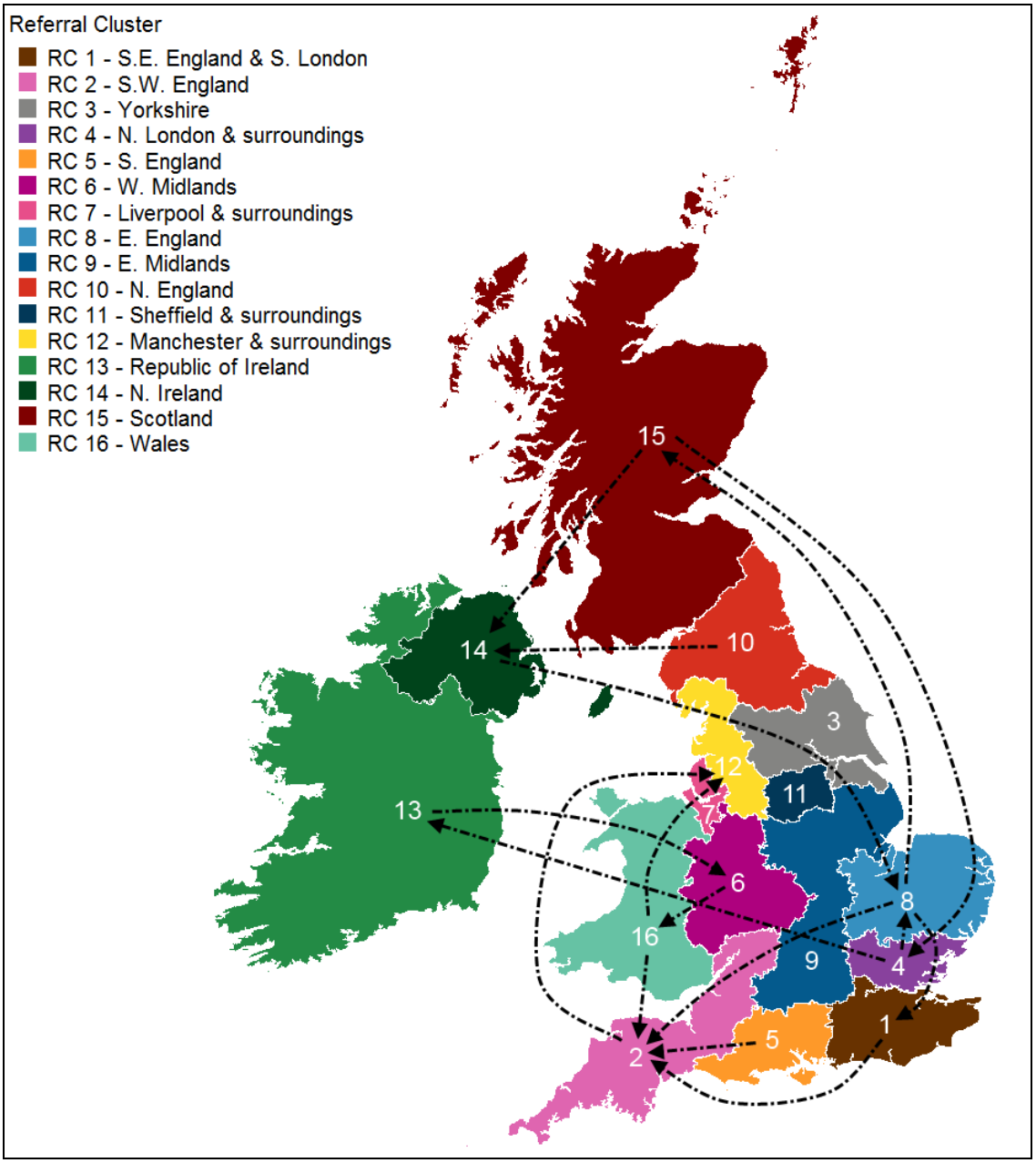


Figure 3.11. Out of 1022 isolates there are 16 Referral Cluster (RC) level Candidate Introductions (CIs) that meet all the conservative thresholds. 72 isolates were initially identified using the TAPO process. These were then reduced to 16 CIs which have only one LSS for the same location that the TAPO method posits as the origin. A large number of these CIs are to non-adjacent RCs. The RC CIs are slightly different to the hospital CIs due to the sub-clade determination on the ML tree. The hospitals have been grouped into 16 separate RCs based on the regions identified in Donker *et al.* (2012, 2014). The hospitals are numbered according to geographic location and RC, with hospitals located in the same RC grouped together.

3.4 Discussion

There are a small number of LSS-containing isolates can be considered as CI events at either the hospital or the RC geographic resolution. This implies that it may be possible to identify introduction events based on a single SNP. Since the considerable majority of isolates have at least one LSS, at either the hospital or RC geographic resolution, this could be a quick way of identifying the origin of any isolate. However, it must be noted that of the 1022 isolates only a few could be considered possible transmission events (127 isolates at the hospital resolution and 72 isolates at the RC resolution), and of these few a tiny fraction survived the strict elimination process described in this chapter (18 isolates at the hospital resolution and 16 isolates at the RC resolution). Therefore, although this method works, it does not have a sufficiently high success rate to be a viable strategy. However, it serves as a proof-of-principle that SNPs can be used to identify transmission events.

A majority of the CIs are between geographically non-adjacent RCs. This is in contrast to the historical finding that ST22 started in the East Midlands and spread across the UK over the next decade (Holden *et al*., 2013). However, a study by Donker *et al.* (2010) implied that the modern day spread of MRSA does not conform to a diffusion pattern, but rather shows a large number of long-range transmission events. Although long-range transmission may seem counter-intuitive, if patient referrals are a method of MRSA transmission (Donker *et al*., 2012; 2014) these results appear reasonable. These long range transmission events could be due to patients being referred to specialist hospitals for specific treatments, or people moving or travelling. Furthermore, long range transmissions might be more likely to be observed since the focal isolate would be from a sub-population with a different complement of SNPs and thus stand out more from the genetic background. Unfortunately these longer range transmissions greatly complicate the possible mathematical modelling of the spread of MRSA strains.

Although the majority of isolates contain LSSs it would be a mistake to assume that the LSS location must be the origin location of the isolate. The LSS definition is dependent on the dataset being used, and therefore insufficient uniform sampling may create incorrect LSSs which could lead to an erroneous assumption of the origin of an isolate. Furthermore, a number of the isolates exhibit multiple conflicting locations, which could be due to insufficient sampling or horizontal gene transfer. Therefore, a way to assess the viability of the LSS location as the origin location is required. Although not all SNPs are LSSs, it is clear that not all SNPs are equally common, either in absolute numbers of isolates harbouring a SNP or in number of

geographic locations. This variable rarity could be informative to the validity of the LSS location as the origin location if the more common SNPs are also found in similar locations. Even if an isolate's exact origin could not be specified, using increasingly rare SNPs would enable the narrowing of the possibilities. It is important to note that the possible origin of the isolate is always within the confines of the sampling, and therefore isolates which may have originated in un-sampled locations or other countries may not be able to be assigned an origin.

The method presented here appears promising, but there are some limitations. Firstly, a phylogenetic tree is required. For small datasets this is not an issue. However, with large and expanding datasets the creation of a phylogenetic tree becomes a lengthy and computationally prohibitive process (Day, 1987). Related to that, the determination of the phylogenetic sub-clades is subjective and requires specialised knowledge. Since the potential CIs are initially determined by their position within a sub-clade, this subjectivity may cause some CIs to be overlooked. Secondly, the criteria thresholds imposed upon the potential CIs are somewhat arbitrary. These thresholds could be modified to be more or less conservative. For example, in this chapter only those sub-clades with five or more isolates were considered. The potential pool of Cis could be increased by reducing this threshold and consider sub-clades with fewer isolates. Finally, those isolates sampled at the beginning of the collection are more likely to present LSSs, since not many isolates have been sampled prior to them. This could artificially inflate the number of LSSs found in the earlier isolates. However, in the Bi-allelic Dataset no trend was found between the number of LSSs an isolate exhibits and the sampling date.

Currently, a manual interpretation of a phylogenetic tree is still required to cross-reference the process of identifying CIs based on a single signature SNP. Therefore, the next step would be to incorporate all the SNPs that an isolate contains to determine the possible geographic origin without using a phylogenetic tree. This process would need to be a broader approach to be able to identify any isolate's geographic origin, within the sampling confines. Furthermore, it might be possible to automate this process, removing the requirement for a manual interpretation of a phylogenetic tree. This could enable the rapid identification of possible transmission events, which could assist in limiting the spread of the pathogen.

3.5 Conclusion

Using the Bi-allelic Dataset, which comprises of non-singleton bi-allelic SNPs, I have shown that it is possible to identify isolates which are introduction events based on a single signature SNP; a Location Specific SNP (LSS). This process currently requires interpretation of a phylogenetic tree and subjective sub-clade definition, and resulted in only a few, very specific, Candidate Introduction (CI) events. Therefore, a broader approach is required. The next step would be to devise a way to identify the possible geographic origin of any isolate within the Bi-allelic Dataset, regardless of LSS incidence and without manual interpretation from a phylogenetic tree.

A novel SNP-based method for determining the origin of MRSA isolates & the identification of transmission events

4

4.1 Background

In this thesis thus far it has been found that there is genetic similarity between MRSA sub-populations due to geographic proximity and level of patient transfer (Chapter 2). It was also found that is sometimes possible to identify introduction events from a phylogenetic tree and characterise them with a signature Single Nucleotide Polymorphism (SNP) for a particular location; a Location Specific SNP (LSS; Chapter 3). In this chapter I will take the identification of introduction events one step further and attempt to remove the necessity of a phylogenetic tree. Furthermore, I will attempt to automate the process of identifying the geographic origin of any given isolate within the confines of the sampling.

Information on the geographic origin of an isolate would enable inform if a particular outbreak is developing into an epidemic, or as found in Köser *et al* (2012), if there is a previously undetected transmission event. As mentioned previously in this thesis, the traditional approach to determine an isolate's origin is to build a phylogenetic tree of all the isolates in question and then look at the phylogenetic clustering (for example, Deurenberg *et al*., 2005; Harris *et al*., 2010; Harrison *et al*., 2013; Holden *et al*., 2013; McAdam *et al*., 2012). This process explained in greater detail in Section 3.2.1.

However, this tree-based approach becomes cumbersome when dealing with large datasets since the computational time required will increase dramatically. Kuhner & Felsenstein (1994) showed that for an increase of 5 taxa the Neighbour Joining (NJ) method led to a 2.5 times longer computation calculation, while Maximum Likelihood (ML) method took 5 times longer. Furthermore, larger trees would take longer to successfully manually identify all sub-clades. The large quantity of data can be difficult to interpret, and there is a drive towards simpler, potentially automated, data interpretation (Köser *et al*., 2012).

The plummeting cost of Whole Genome Sequencing (WGS), currently less than £50 for one MRSA isolate (Priest *et al*., 2012), opens up opportunities for the use of WGS as a routine practice in healthcare institutions in the near future. If every case of MRSA bacteraemia identified in a healthcare institution was sequenced using WGS, a large repository of MRSA

genomes would soon be built up. If this sequencing practice is adopted across the UK and Ireland each institution could add their genomes to a collated online database. Therefore, a computationally cheap method could be integrated into the analysis pipeline of the MRSA isolates in each healthcare institution which would use the collated online database to determine the possible geographic origin of the MRSA isolate.

The ML phylogenetic tree in Chapter 2 (Figure 2.4) was used to determine the possible geographic origin location of the isolates. However, if a different phylogenetic construction method was implemented (e.g. Neighbour Joining) then a slightly different set of Candidate Introduction (CI) events may have arisen. Finally, if a different person was to characterise the same phylogenetic tree there may be discrepancies between how each person defines the sub-clades. Since it is this sub-clade definition which is used to define the geographic origin of an isolate this discrepancy would also result in a different set of CIs. Therefore, an automatic method is required that would consistently provide the same geographic origin for a given isolate. One potential approach would be to examine an isolate's SNPs and use these SNPs to posit a potential geographic origin. This would obviate the need for constructing a phylogenetic tree and remove some of the subjectivity of the current standard process. Finally, by-passing the phylogenetic approach would reduce the time required in determining the geographic origin for an isolate.

In this chapter I present a novel method which will examine the SNPs an MRSA isolate harbours to determine which geographic location the isolate has originated from by comparison to the SNPs present in the other MRSA genomes. The identification of the SNPs used in this is described in Section 2.2. I will attempt to elucidate the possible geographic origin of the isolate at both the hospital and the RC geographic resolution. RCs are regions within which hospitals refer patients to one another and are described in Donker *et al.* (2012, 2014) and Section 2.3. This method will attempt to circumvent the traditional tree based approaches and may be integrated into the routine pipeline of MRSA diagnosis in hospitals to provide early warning of a new import. I start by explaining, with a hypothetical example, the process by which SNPs are used to determine an isolate's origin. I will be using the Bi-allelic Dataset as in the previous chapter(s); 5469 non-singleton bi-allelic SNPs which are present in at least 2 of the 1022 isolates of EMRSA-15 in Clonal Complex 22 present across 46 hospitals in the UK and Ireland (UK&I) between 2001 and 2010. Therefore, any posited origin for an isolate will be constrained to the locations sampled. I will use two sub-sets of isolates from the Bi-allelic Dataset to test this novel method, and an additional set of isolates sampled from 2011 and 2012. The first group of test isolates will contain the Candidate Introductions (CIs)

determined in Chapter 3, which might be expected to show a clear signal of transmission. This group of isolates is termed the CI Test Subset. The second group of test isolates will contain all the isolates sampled in 2010. The genetic profiles of these isolates are very varied and there are no prior preconceptions of the geographic origin of these isolates. This group of isolates is termed the 2010 Test Subset. Using these two subsets of isolates I show that it is possible to determine the potential geographic origin location, within the confines of the sampling, of an isolate based on the SNPs it contains. I also investigated how the computational time and efficiency of SnAPO will scale with an increasing database size. I show that SnAPO would likely be computationally faster than standard phylogenetic methods with ever increasing database sizes.

Additionally, the BSAC isolates sampled in 2011 and 2012 were used to test how SnAPO performs if the sampling location metadata is obscured. These isolates were processed using SnAPO and investigated if the origin location could be determined without knowing the sampling location. Finally, in a test of SnAPO's robustness and ability to cope with data from other studies, isolates obtained in an alternate study were examined and processed using SnAPO. The isolates were from Holden *et al* (2013), which examined the spread of EMRSA-15 across the world and used phylogenetic analysis to show the origin of this strain to be in the English Midlands (see Section 1.4.2 for more details). This study contains a large number of isolates from all across the world and so would provide a diverse dataset which could be used to determine the various types of output that may be obtained through SnAPO for a wide range of isolates.

4.2 Method

4.2.1 A SNP-based assignment of pathogen origin

There is high variability in the rarity of the SNPs in the Bi-allelic Dataset; some are found in many isolates, while others are harboured in a select few (Section 2.4). In Chapter 3 I demonstrated that certain SNPs can be used as a signature for a specific geographic location (Location Specific SNP; LSS), and this signature can be used to identify some Candidate Introduction (CI) events. However, it was found that only a few CIs can be identified in this way. Therefore an approach with broader application might be necessary. The method described in this section could be applicable to all isolates.

It is assumed that any SNP shared between isolates is an indication of a shared ancestor of those isolates. This assumption can be made due to the clonal nature of MRSA that results in a very low rate of homoplasy and recombination in the core genome (Castillo-Ramírez *et al*., 2011). Furthermore, unless specifically selected against or lost through genetic drift, it is known that SNPs accumulate in a genome over time and can be inherited vertically (Thomas *et al*., 2011). Since I am attempting to elucidate the potential geographic origin of an isolate I will examine the geographic incidence of the individual SNPs. It must be noted that this is a heuristic method, however in Chapter 5 a Bayesian analysis is applied to this problem with similar results.

A hypothetical example, with typical numbers and distributions of SNPs, will be first used to illustrate this method. In the results Section 4.3.1 three real examples are provided as case-studies to further illustrate this method. The first step in this heuristic method was to choose a focal isolate. Any isolates sampled after the focal isolate were excluded from the analysis. Once this was done the SNPs the focal isolate harbours were determined and, for each SNP (*S*), how many isolates (*n*) and locations (*L*) they have appeared in before. This information can be presented as a bar-plot for each SNP (Figure 4.1).

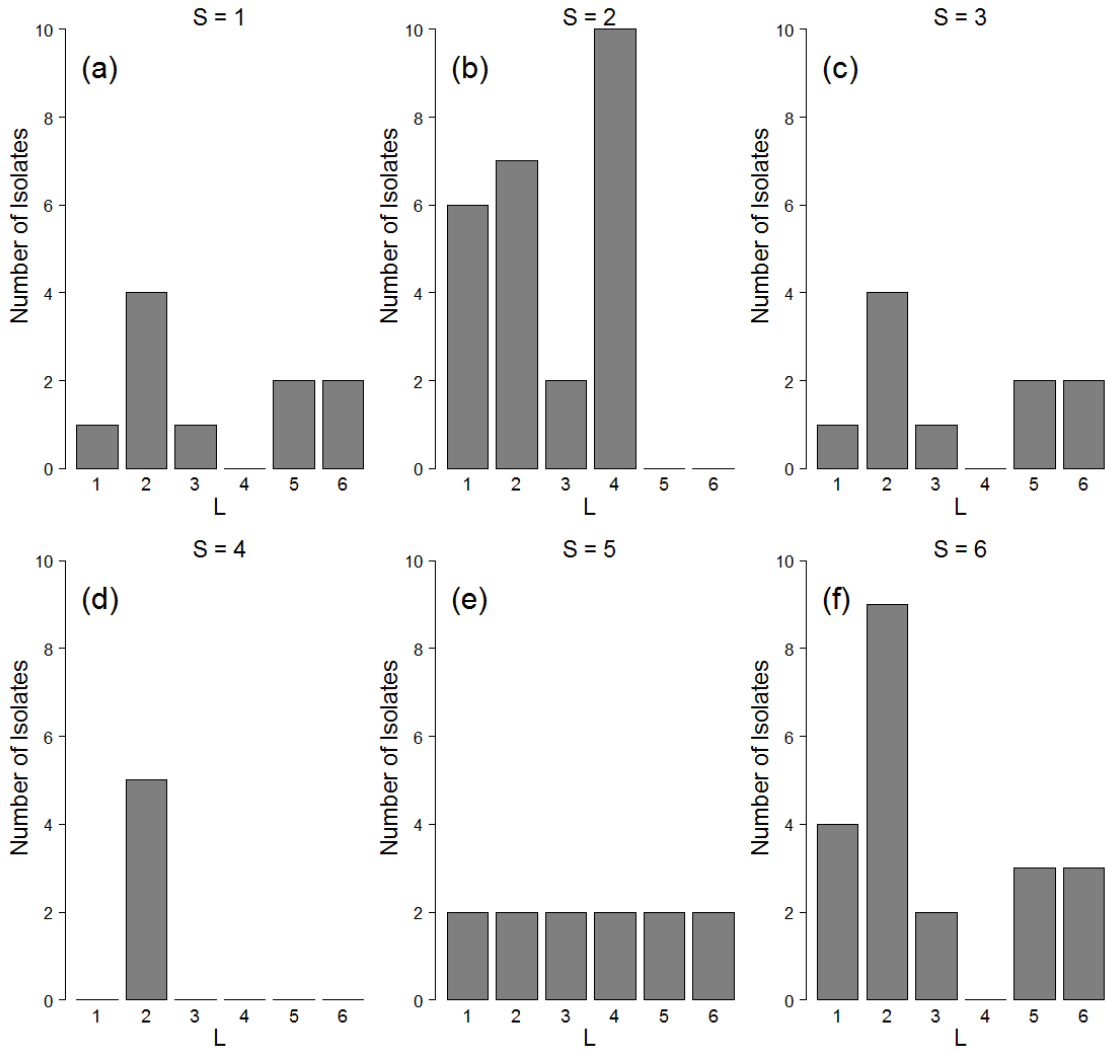


Figure 4.1. Each SNP is assigned an index number (*S*). The number of isolates (*n*) each SNP is found in for each location (*L*) is obtained; *n(L,S)*. In this hypothetical example I provide examples of typical SNP geographic incidence distributions. In (a) the majority of isolates with S = 1 are in L = 2; in (b) the SNP appears in many isolates; in (c) there is the same distribution as in (a); (d) is an LSS for L = 2; (e) is a SNP that is found in every location equally; (f) is a SNP which shows a similar distribution to (a) and (c), except in a few more isolates. These SNP geographic distributions are conflicting in their indication of the origin of the hypothetical isolate. Although there is an LSS for L = 2 in (d), there are many isolates from L = 4 that share S = 2 in (b). S = 2 is likely an older SNP that has penetrated more isolates, while S = 4 is probably a younger one.

The incidence of each SNP in the hypothetical focal isolate has been identified (Figure 4.1). In this hypothetical example there is an LSS for Location 2 (S = 4, Figure 4.1d) indicating that the isolate in question might have originated, within the confines of our sampling, from Location 2. However, there is some conflicting evidence that Location 4 might be the origin, since many isolates in that location share the SNP S = 2 (Figure 4.1b). Therefore, it should not be immediately concluded that the location of the LSS is the geographic origin of the focal

isolate. If the location of the LSS is indeed the origin of the focal isolate then this location will be supported by other, more common, SNPs which the focal isolate harbours. A lack of support could be indicate that the LSS might have arisen through homoplasy, recombination or under-sampling.

The more common SNPs (such as S = 2 and S = 6 in Figure 4.1), which are usually older (Section 2.4.2, Figure 2.10), provide less information about the relatedness between isolates than the rarer SNPs. Although it is important for the location of the LSS to have support of more common SNPs it can be assumed that LSSs, as a signature for a specific location, provide greater information of the most recent origin of the focal isolate. Therefore, the next step was to standardise each SNP's distribution but allowing the rarer SNPs (such as LSSs) greater influence on the origin of the focal isolate. For each focal isolate SNP (S = 1 to S = n) the number of other isolates that harbour that SNP in a particular geographic location was standardised by the total number of other isolates that harbour that SNP in all locations. This will be termed this *ṅ*. Therefore, an LSS would have a value of ṅ = 1 for the location of the LSS; a SNP seen in equal numbers of isolates in two locations will have a value of ṅ = 0.5 for each location, and so on. An increase in the geographic dispersal of a SNP will correspond with a decrease in the value of that SNP at any one geographic location. This standardisation allows the comparison of SNPs within the focal isolate (Figure 4.2).

There is variation in the number of samples for each location in the Bi-allelic Dataset (Section 2.3, Figure 2.2b). Therefore, it could be argued that the sampling effort in each location should also be taken into account; for example, by dividing each SNP's incidence at a location by the total number of isolates sampled at that location. However, this will artificially inflate those locations that were under-sampled. Therefore, this factor is not included in the analysis but instead remain aware that the interpretation of the origin of the focal isolate may be influenced by this choice. This will likely remain an issue until routine sequencing of MRSA isolates in all healthcare institutions becomes commonplace.

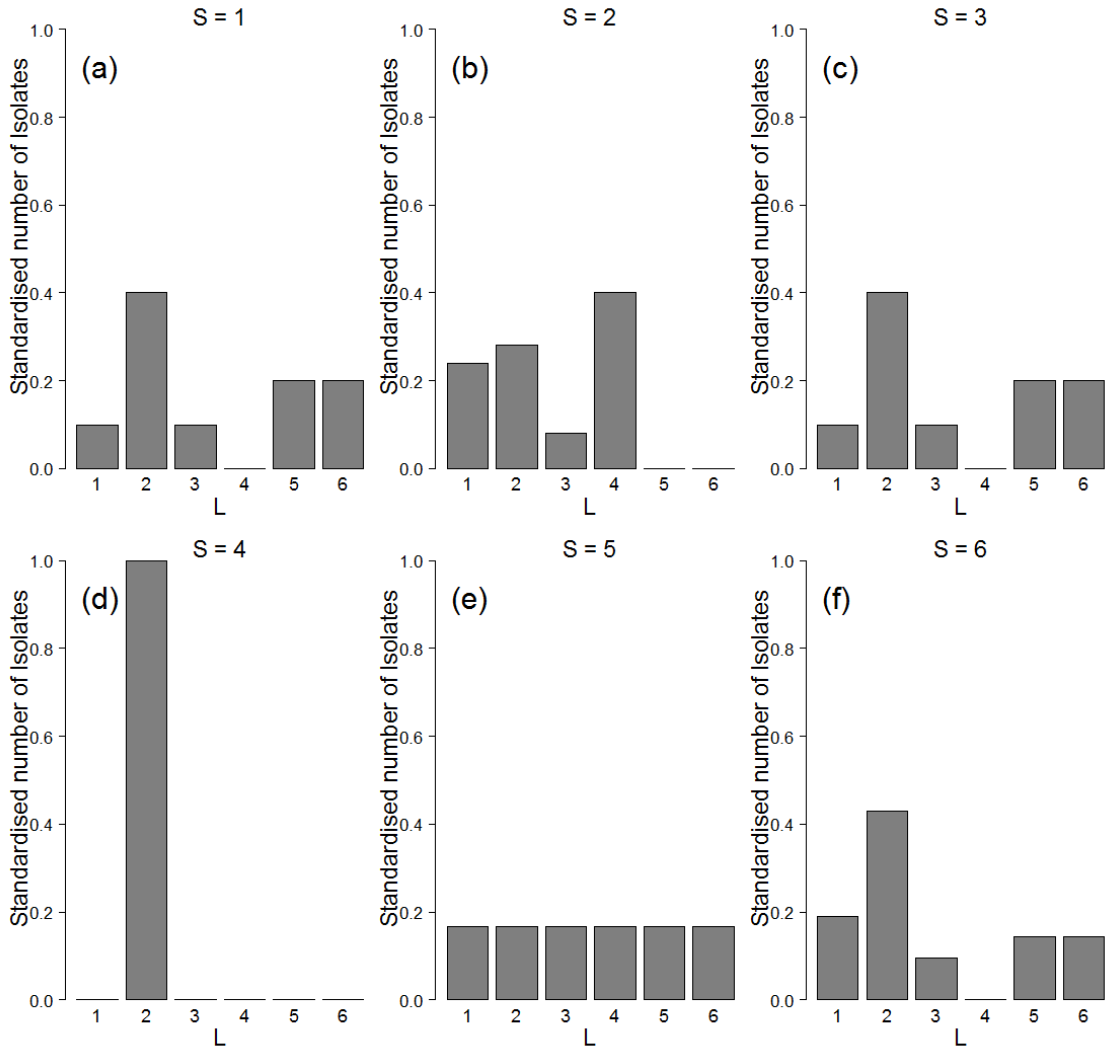


Figure 4.2. The number of isolates for each location for a SNP (*n(L,S)*) is standardised (*ṅ*) by the total number of isolates which harbour the SNP (*∑n(L,S)*); *ṅ = n(L,S)/ ∑n(L,S)*. An LSS, such as S = 4 in (d) gives a standardised value of 1, indicating strong support for that location. The standardised values for more common SNPs provide less informative support, such as S = 2 and S = 5 in (b) and (e). In this hypothetical example, the standardised incidences of the SNPs begin to indicate more strongly that the possible origin is L = 2, but there are other locations with SNPs in common with the focal isolate. Therefore a way to combine the SNPs into one diagnostic value is required.

The standardised incidence of each SNP based on its rarity has been determined (Figure 4.2). There is still some conflicting evidence on the origin of this hypothetical isolate, so all information must be considered before concluding the possible origin. Therefore, the next step will combine the standardised incidence of each SNP to give a total value for each location. However, there are some issues that need to be addressed first. Thus far it has been assumed that each SNP is an individual piece of information, but in this hypothetical example S = 1 and S = 3 have identical incidences (i.e. in every isolate that S = 1 is present, S = 3 is also

present). SNPs with identical incidences were defined as SLSs in Section 2.4.3. It is possible that combining these SNPs to give a total value for each location might count one piece of information twice. If it is decided that each SNP which shows identical distribution is only one piece of original information there is a secondary problem. In this hypothetical example S = 6 has very similar distribution as S = 1 and S = 3. Although S = 6 is present in more isolates, every isolate which harbours S = 6 also harbours S = 1 and S = 3. It could be said that S = 1 and S = 3 are "nested" within S = 6. It now becomes difficult to decide how to treat these nested SNPs, especially since there are situations where a SNP is not fully nested within another. Therefore, although there may be some issues with counting SNPs twice, the simplest case is to continue to assume that each SNP is an independent piece of information. Furthermore, this is in line with the assumption that each SNP is indicative of a common ancestry. However, the issue of identical and nested SNP incidences is certainly an avenue of investigation waiting to be resolved.

Therefore, the standardised incidences of all SNPs at each location were summed *(ṅ)* to give a total value for each location (*p*) (Figure 4.3a). These values were standardised to facilitate comparison between isolates by dividing each location's value by the total number of SNPs the target isolate harbours. This final output can be presented as either a diagnostic proportion or percentage of origin (Figure 4.3b). This is termed the Diagnostic Origin Value (DOV). A higher DOV indicates that the particular location is the potential geographic origin of the focal isolate. However, it is important to note that unless there is a location with an obviously highest DOV, the interpretation of this output could be subjective. However, for a given dataset and a given focal isolate SnAPO would always achieve the same DOV result.

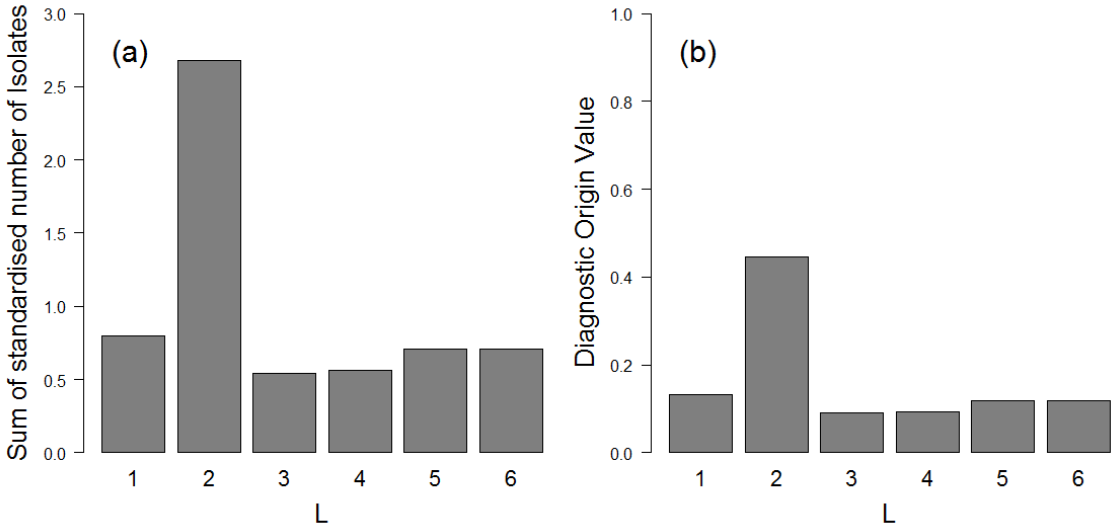


Figure 4.3. Each location is given a total value (*p*) which is the sum of each SNP's proportion at that location (*ṅ*) (a). In order to compare the results for different isolates each location value is standardised by the number of SNPs (*s*) in the focal isolate (b). This results in a Diagnostic Origin Value (DOV) for each location; *DOV = p/s.* In this hypothetical example, the final standardisation (b) gives a DOV maximum value above 0.453 for L = 2, and the next highest (0.178 for L = 1) is considerably lower. Therefore, L = 2 is the possible origin location for this hypothetical isolate within the confines of the sampling.

In summary, the DOV for a given geographic location for a target isolate was obtained by summing together all the standardised instances of each SNP at that location prior to the sampling of the target isolate, and then dividing by the total number of SNPs in the target isolate (Equation 4.1). This heuristic method is termed the SNP-based Assignment of Pathogen Origin (SnAPO):

$$DOV(L) = \frac{\sum^{n(L,S)}/_{\dot{n}(S)}}{s} \qquad \textbf{Equation 4.1}$$

Where *n(L,S)* is the number of isolates at location *L* harbouring SNP *S*, *ṅ(S)* is the number of isolates that harbour the SNP *S* in all locations, and *s* is the number of SNPs harboured in the target isolate.

4.2.2 Processing the CI Test Subset with SnAPO

To test the validity of this heuristic method the posited origin locations determined from SnAPO was compared with those determined using the Tree-based Assignment of Pathogen Origin (TAPO; as presented in Chapter 3). With TAPO there were 127 isolates that could be considered as potential Candidate Introductions (CIs) on the Maximum Likelihood (ML) tree, at the hospital level of geographic resolution. These isolates came from sub-clades

of varying size but all had relatively high confidence with regards to bootstrap values. In Chapter 3 the 127 hospital Cis were reduced down to only 18, based on strict criteria. However in this chapter all 127 isolates in the CI Test Subset were processed using SnAPO and obtained a diagnostic value for the hospital origin of each of the isolates. Since these isolates were identified by the phylogenetic tree as possible introduction events it might be expected that there is some more obvious geographic signal than in isolates which are not CIs. However, the CIs show a range of SNP incidence, with some of the 127 CIs exhibiting LSSs for multiple locations and some not exhibiting any LSSs. Processing the CI Test Subset isolates with SnAPO would test whether the origin location predicted by SnAPO concurs to the location posited by TAPO. Those isolates which only harbour very common SNPs might prove to be difficult to obtain a clear origination signal.

4.2.3 Processing the 2010 Test Subset with SnAPO

Thus far, SnAPO has only been used on isolates which are expected to have an origination signal, based on their appearance on the ML phylogenetic tree. The next step was to determine if SnAPO is a viable analysis approach for any isolate. SnAPO was tested on the 90 isolates sampled from 2010. Although a few of these 90 test isolates are in fact CIs, there is considerable variation in their phylogenetic profile. Some are isolates found in sub-clades that are too varied or too small to determine geographic origin via TAPO, other isolates were unable to be assigned to a sub-clade, and finally some of the isolates are predicted to have originated from their sampling location.

The 90 isolates were processed using SnAPO, ensuring the temporal order of sampling was taken into account. Three colleagues were enlisted to help determine, individually and separately, where they predict the 90 test isolates may have originated from, based on the ML phylogenetic tree (Chapter 2, Figure 2.4). The investigators were asked to also indicate whether they believe the isolate came from the predicted location with high or low confidence. The investigators were allowed to use any preferred method to determine origin location and confidence. Once the investigators had completed the task they described the process they used. Although they all operated independently and therefore had slighty differing methods there was a consistent base methodology which I will summarise and paraphrase as follows. If an isolate was grouped into one clade with isolates all from one location then that was the origin location; if a target isolate was grouped between two clades each containing only isolates from one location then the closest clade location was the origin; if a target isolate was grouped with isolates from multiple locations then the investigators used

their best judgement. The confidence of the prediction was also based on the investigators best judgement. The predictions of the three investigators were compared to each other and to the predictions obtained by SnAPO. Not only was the location that was predicted comapred, but also the confidence of the prediction when compared to the numerical DOV.

It is important to note here, that this is a slight change in the Bi-allelic Dataset. Prior to this all 1022 isolates were used, with 5469 non-singleton bi-allelic SNPs. For this particular test the 90 isolates sampled in 2010 (the 2010 Test Subset) are compared with all the ones sampled previously. The 931 isolates sampled between 2001 and 2009 is termed the Comparison Subset. All 5469 SNP positions are retained, even for those SNP positions where the SNP is only found in isolates in 2010. Furthermore, after processing each isolate from 2010 it will then be added to the Comparison Subset. Therefore, the first isolate in 2010 will be compared to the 931 isolates sampled before it, while the last isolate in 2010 will be compared to the 1021 isolates which have been previously sampled and added to the Comparison Dataset. This will mimic the potential future practical application of the method as part of the routine screening procedure in a healthcare institution.

4.2.4 Determining the speed of SnAPO

The 2010 Test Subset was used to determine the computational speed of SnAPO with increasing database size. The time required for SnAPO to process each of the 90 test isolates when using the full Comparison Subset (n = 932) and then the time required to process each isolate when using half the Comparison Subset (n = 466) was calculated. This will give an idea of what effect doubling the size of the dataset will have on processing time. As previously stated after each test isolate is processed it is added to the dataset before moving on to the next test isolate.

4.2.5 Processing the 2011 and 2012 isolates

The 17 isolates sampled in 2011 and 2012 were used to determine how SnAPO would perform when the sampling location was unknown. To this end the metadata indicating sampling location was removed and the isolates were processed without any knowledge of where each isolate was sampled from. SnAPO was used to determine the possible origin location for each of the 2011 and 2012 isolates and then this prediction was compared against the retrieved sampling location. After each isolate was processed it was added on to the dataset (as described in the previous section). This process will mimic a real world situation were the sampling location of an isolate is unknown.

4.2.6 Processing isolates from Holden *et al* (2013)

There are 193 isolates analysed in Holden *et al* (2013) and, with the permission of the author, a subset of these isolates was used to test SnAPO's performance on isolates that come from a different dataset. The isolates were sampled from 14 different countries, between 1991 and 2009 (Figure 4.4). Of the 193 isolates, 86 were sampled from the UK. However, the sampling locations within the UK are, in the main, different from the ones used in the data previously presented in this thesis. Only two isolates (X07_1361_K, X07_2789_T) in 2006 and 2007 were sampled from locations that were also sampled in the Bi-allelic Dataset.

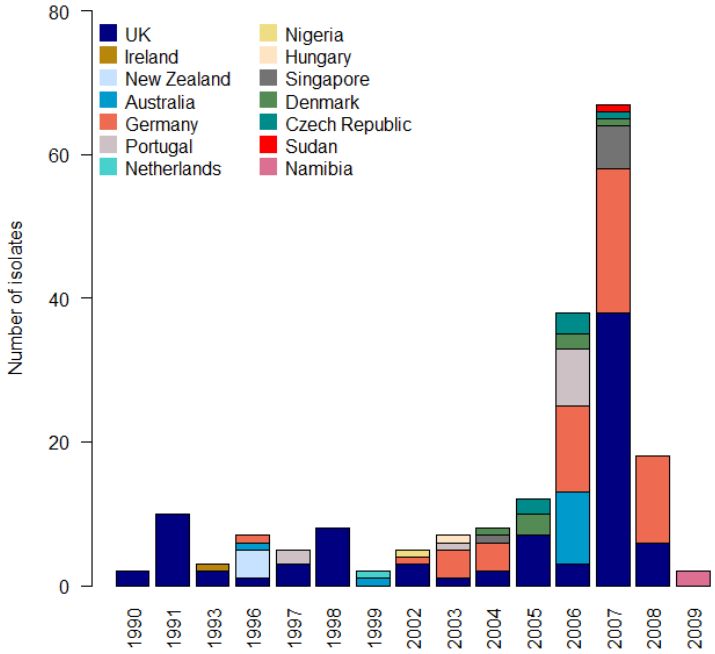


Figure 4.4. The sampling effort for the Holden *et al* (2013) study, including MSSA isolates. Of the 193 isolates in this study, 86 isolates were sampled from various locations within the UK, while the rest were from countries around the world. More than half of the isolates were sampled in 2006 and 2007 (n = 38 and n = 67, respectively).

SnAPO was used to process the isolates, excluding MSSA, sampled in 2006 (n = 35) and 2007 (n = 62) in a similar way to that described previously in this chapter. The isolates in the Holden *et al* (2013) study were mapped to the same reference genome as the ones in the Bi-allelic Dataset (see Section 2.2) and so the SNPs would be comparable. However, any SNPs not found in the Bi-allelic Dataset were excluded, which left 931 SNP positions that appear in the isolates from Holden *et al* (2013) and the Bi-allelic Dataset. As described previously, the isolate was compared to isolates which have been sampled prior to the target isolate. However, the isolate was not added to the database once it was processed. All except 2 of the isolates from 2006 and 2007 were sampled from countries other than the UK, or from locations within the UK that have not been represented in this thesis. Therefore, there might be ambiguous signals

for all those isolates which originate from locations which do not appear in the Bi-allelic Dataset.

The isolates sampled from 1991 (n = 10) were also processed with SnAPO. These isolates were sampled before any isolate in my database, therefore these isolates were compared to the entire Bi-allelic Dataset. This was done to test how SnAPO processes isolates that may be incorrectly dated. The isolates sampled in 1991 would likely have inconclusive and ambiguous SnAPO output, with no clear origin location.

Finally, a single location was chosen that was most heavily sampled between 2006 and 2007 in the Holden *et al* (2013) study and processed the isolates from this location again, but this time each target isolate was added on to the dataset once it was processed with SnAPO. The location used was London (UK), which was sampled 14 times in 2007. By adding on each isolate an increasing signal for London might be expected to be observed in the subsequent isolates. However, the location specified by Holden *et al* (2013) was not more specific than the city the isolate was sampled from and so it is possible that these 14 isolates come from different hospitals within London.

As previously stated a maximum DOV of higher than 40% will be used to be indicative of a possible origin location according to SnAPO. However, I reiterate that this is an arbitrary value and other characteristics of the SnAPO output should be considered; for example, the high DOV values for other locations and the difference between these values.

4.3 Results

I initially present (Section 4.3.1) three isolates that have been processed using SnAPO to illustrate and further explain the SnAPO process. Section 4.3.2 will investigate the output of SnAPO on the Candidate Introduction events (CIs) identified in Chapter 3, to determine if SnAPO can predict the same posited geographic origin as the Tree-based Assignment of Pathogen Origin (TAPO) method. I will investigate if SnAPO can predict the same posited geographic origin as TAPO for any of the isolates sampled in 2010 (Section 4.3.3). The speed of SnAPO is also described (Section 4.2.4). Finally I investigate the isolates sampled in 2011 and 2012 (Section 4.2.5) and those sampled in Holden *et al* (2013) (Section 4.2.6).

4.3.1 Case studies of the SNP-based assignment of pathogen origin process

The three isolates used as examples were all sampled in 2010 and were all chosen from the 127 potential CIs. The examples that were chosen from the 127 CIs are those isolates that might be expected to have some signal for a geographic origin, as determined by the Maximum Likelihood (ML) phylogenetic tree. Initially, the ML phylogenetic profile of each isolate was analysed to determine the possible geographic origin based on a Tree-based Assignment of Pathogen Origin (TAPO; see Section 3.2.1). The isolates were processed using SnAPO in an attempt to elucidate the possible geographic origin of the isolate at both the hospital and the RC geographic resolution. The possible hospital origin is displayed as a bar chart, while the RC origin is shown as a geographic heat-map with the appropriate RC geographic divisions.

I first present an isolate where the SnAPO output indicates a very strong connection to one particular hospital. The second case study isolate is one where SnAPO indicates the origin to be within a single RC but not a single hospital. The last case study isolate presented is ambiguous and could be considered to have originated from multiple geographic locations, or from a location that was not sampled.

4.3.1.1 Case Study 1 – Isolate X7564_8.37

Isolate X7564_8.37 was sampled from Chelsea in 2010. This particular isolate harbours 62 SNPs; 25 of these SNPs are LSSs for London St. Mary. The ML phylogenetic tree places it in a sub-clade of London St. Mary isolates (Figure 4.5a), so it could be posited that London St. Mary is the origin of this particular isolate. SnAPO provides an output of DOVs for each sampled hospital (Figure 4.5b). In this particular case there is a very clear signal that this isolate has originated from London St. Mary, as predicted by the ML tree. The RC level output can be

displayed as a geographic heat-map of the UK and Republic of Ireland (Figure 4.5c). There is a clear signal for the South London and environs RC. Therefore, Isolate X7564_8.37 is a clear example of an isolate originating from one particular RC (South East England and South London) and one particular hospital (London St. Mary).

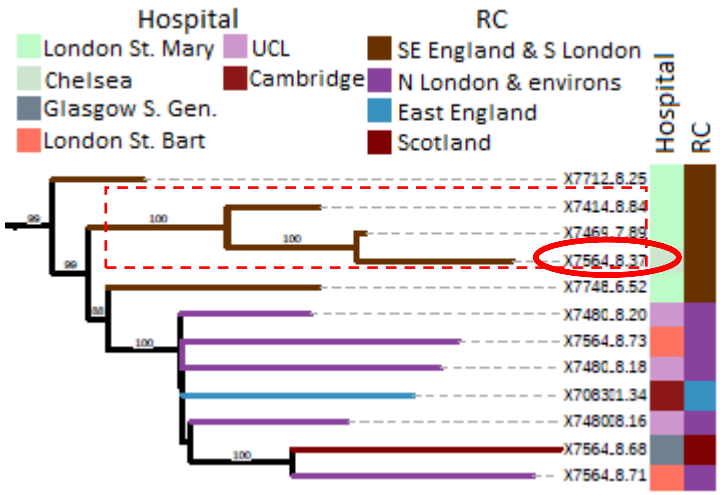


Figure 4.5a. Part of the ML phylogenetic tree depicting the sub-clade (red dashed box) containing isolate X7564_8.37 (red circle). Branch length is proportional to difference in the number of SNPs in the isolates. The numbers above the branches leading to bifurcations are the bootstrap values. If these numbers are absent then this split has a bootstrap value less than 80. The left colour column indicates the hospital the isolate was sampled from, while the right colour column indicates the RC. Branches are also coloured by RC. This sub-clade clustering indicates that the focal isolate could have originated from London St. Mary.

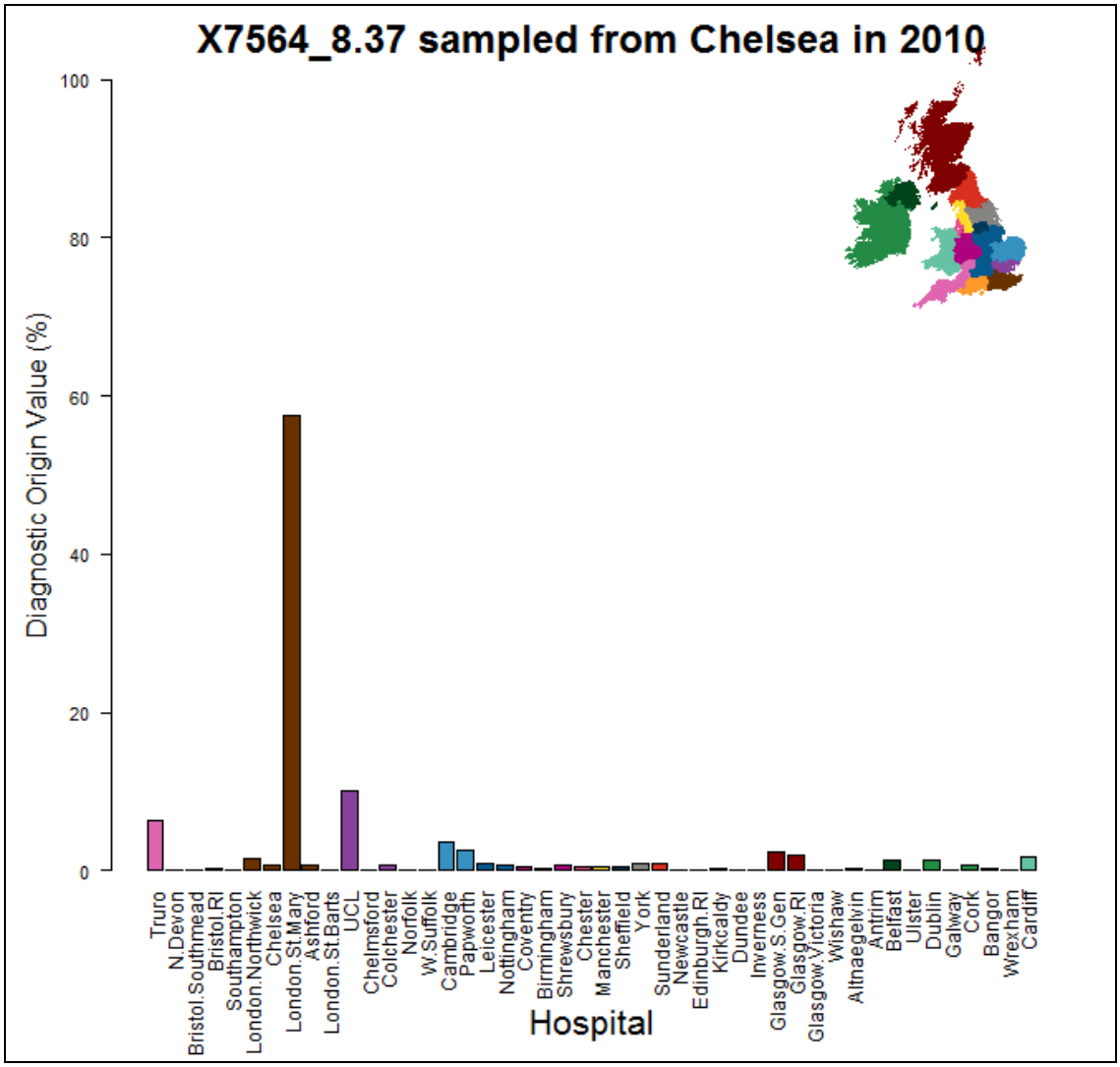


Figure 4.5b. The output for isolate X7564_8.37 shows a clear signal for London St. Mary as the origin hospital. The hospitals along the x-axis are laid out in a geographic order, with each Referral Cluster's hospitals adjacent to each other. The y-axis denotes the DOV, as a percentage. The bars are coloured by the Referral Cluster, which is provided as a graphical legend.

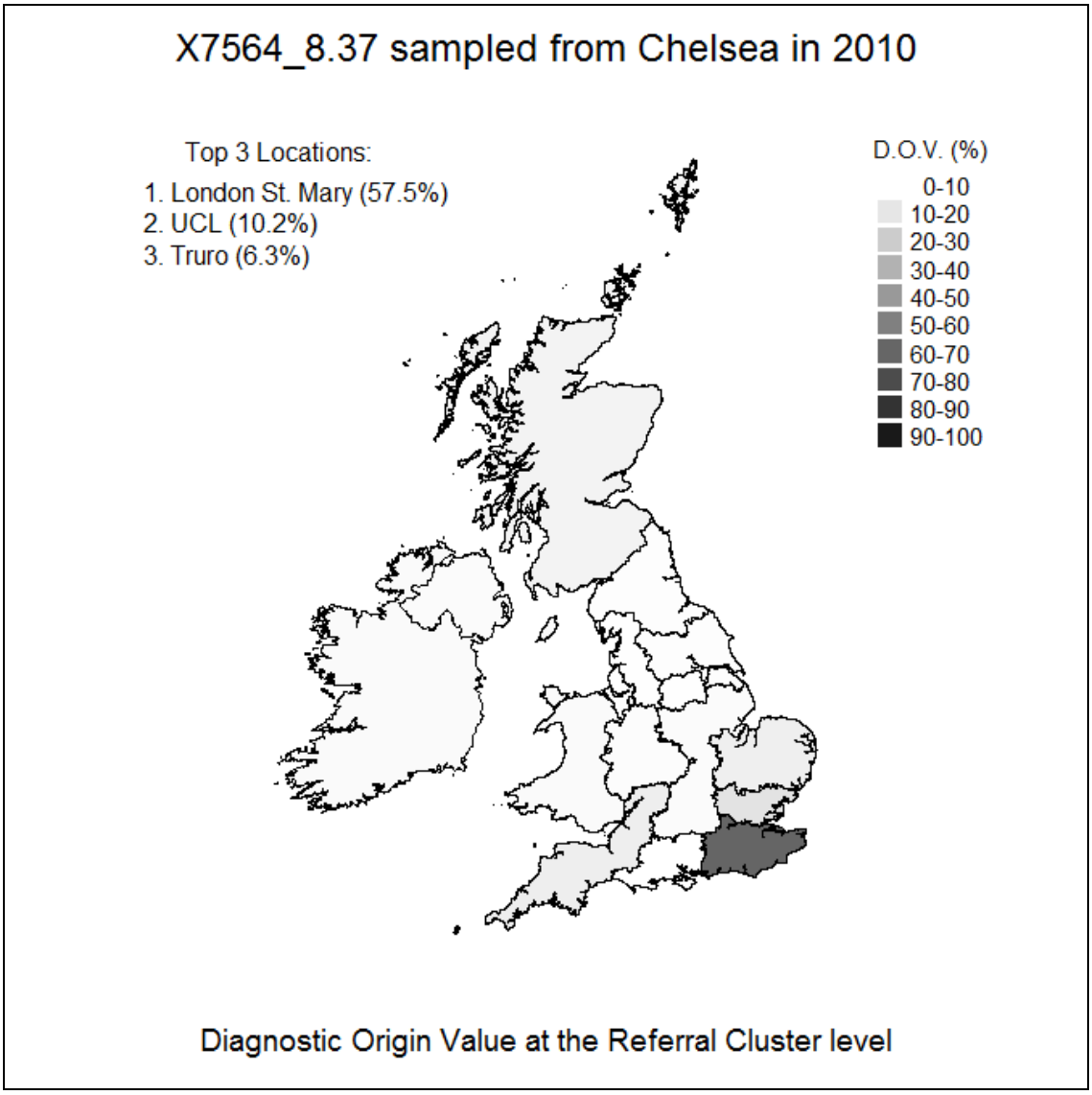


Figure 4.5c. The geographic heat-map of origin for isolate X7564_8.37 denoting specific RCs. There is a clear signal that the isolate originated from the South London and environs RC. Also provided are the top three potential origin hospitals. An increase in opacity indicates a greater DOV for that RC; a key is provided with 10% increments.

4.3.1.2 Case Study 2 – Isolate X7748_6.80

Isolate X7748_6.80 was sampled from Bangor in 2010. The ML phylogenetic tree places the isolate in a sub-clade containing a majority of isolates from West Suffolk (Figure 4.6a), so it is possible that the origin of this isolate is West Suffolk. This isolate harbours 25 SNPs; 1 of these SNPs is an LSS for Cork. This LSS appears to contradict the ML tree clustering, thereby obscuring an obvious origin for this focal isolate. SnAPO provides output DOVs for each sampled hospital (Figure 4.6b). There is a signal that this isolate has originated from West Suffolk, as predicted by the ML tree. However there is still some contention since the signal is

not as clear as in Case Study 1, with no one particular hospital the obvious geographic origin. However, the highest DOVs are all close together, indicating that it might be originating from one particular RC. This can be graphically represented as a geographic heat-map of the UK and Republic of Ireland (Figure 4.6c). Therefore, although in this focal isolate it is not perfectly clear which hospital was the origin, it can be considered that the East of England is the origin RC. Furthermore, within that cluster West Suffolk has the greatest DOV, indicating that this is a possible origin hospital.

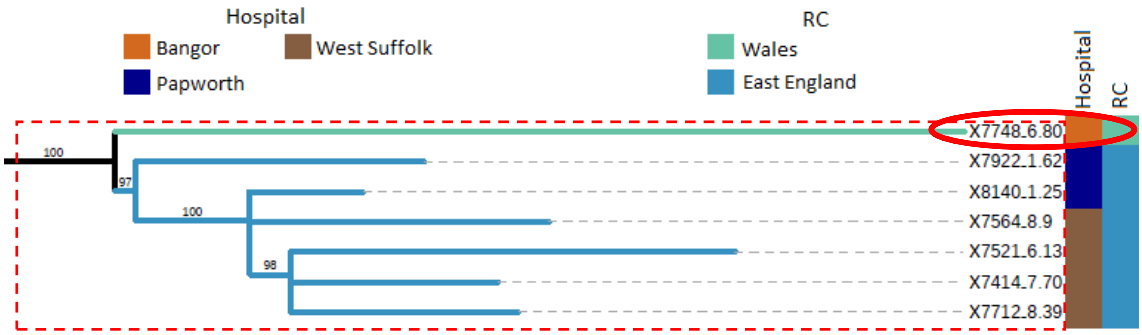


Figure 4.6a. Part of the ML phylogenetic tree depicting the sub-clade (red dashed box) containing isolate X7748_6.80 (red circle). Branch length is proportional to difference in the number of SNPs in the isolates. The numbers above the branches leading to bifurcations are the bootstrap values. If these numbers are absent then this split has a bootstrap value less than 80. The left colour column indicates the hospital the isolate was sampled from (orange = Bangor, brown = West Suffolk, blue = Papworth), while the right colour column indicates the RC (turquois = Wales, light blue = East England). Branches are also coloured by RC. This sub-clade clustering, with the majority of isolates sampled in West Suffolk, could indicate the origin of the focal isolate is West Suffolk.

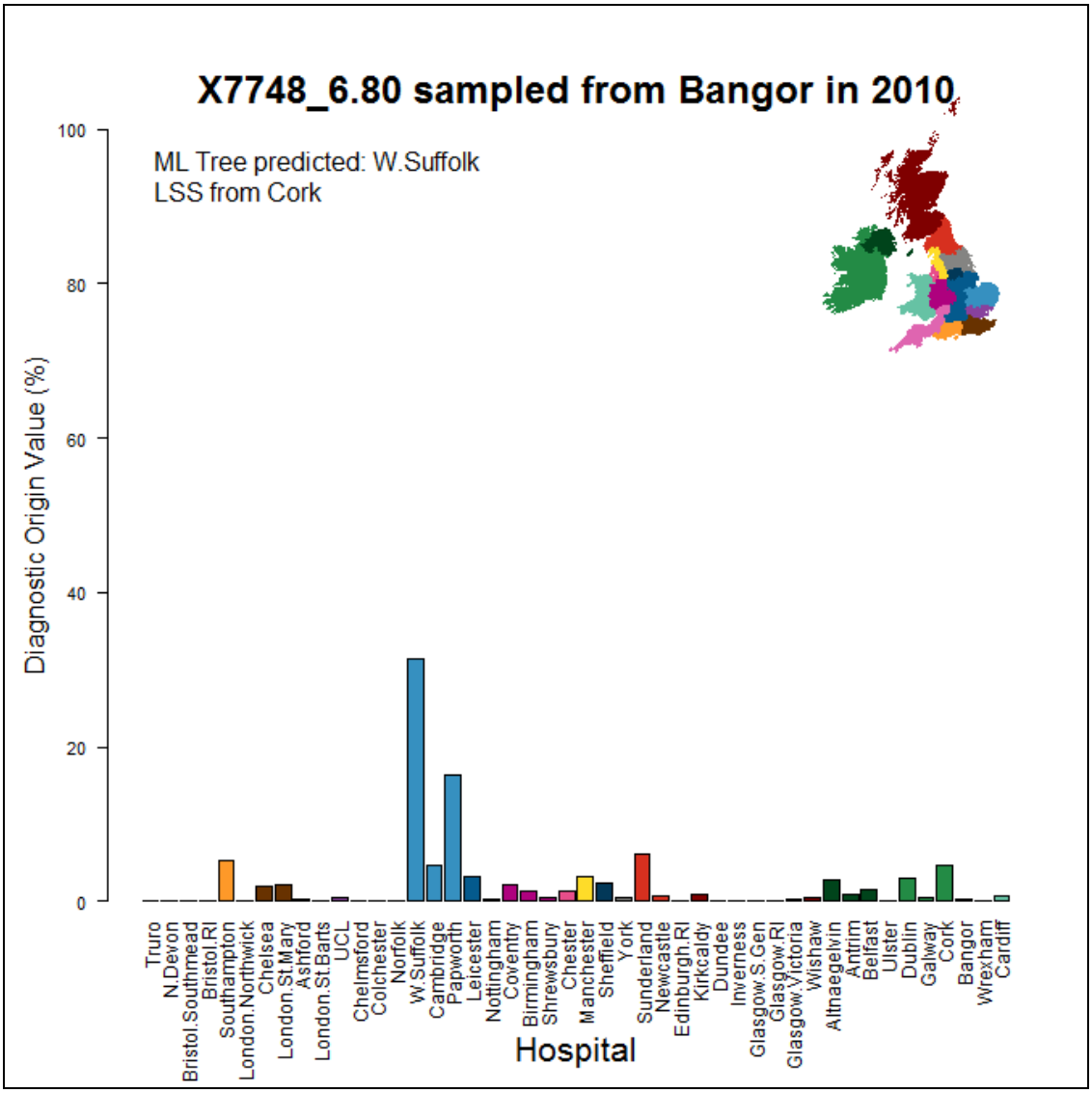


Figure 4.6b. The output for isolate X7748_6.80 shows that there is a signal for West Suffolk as the origin hospital. However, this signal is not very strong and there is a second signal for Papworth, which is in the same RC. The hospitals along the x-axis are laid out in a geographic order, with each Referral Cluster's hospitals adjacent to each other. The y-axis denotes the DOV as a percentage. The bars are coloured by RC, which is provided as a graphical legend.

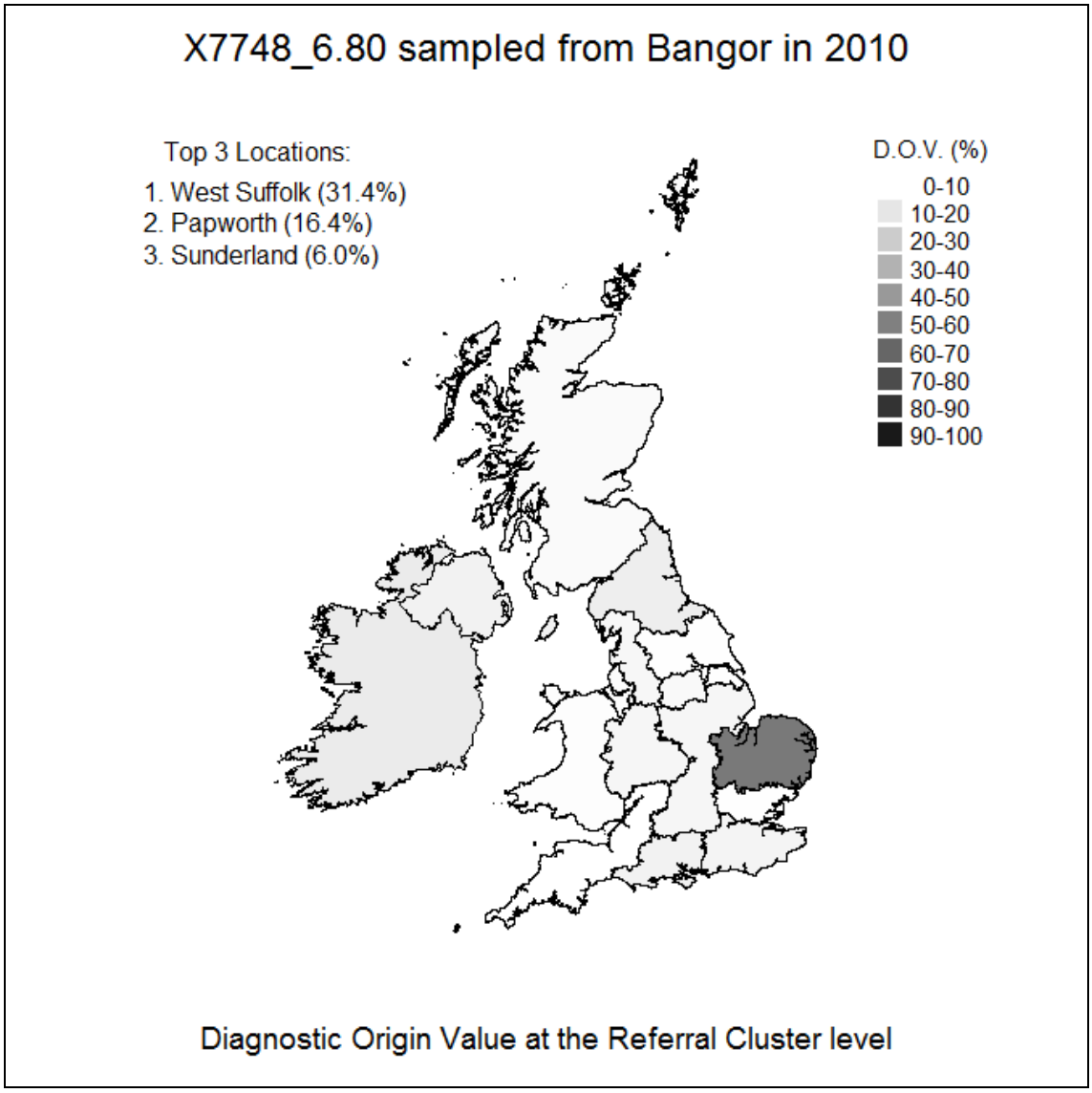


Figure 4.6c. The geographic heat-map of origin for isolate X7748_6.80 denoting specific RCs. There is a clear signal that this isolate originated from the East of England RC. Also provided are the top three potential origin hospitals. An increase in opacity indicates a greater DOV for that RC; a key is provided with 10% increments.

4.3.2.3 Case Study 3 – Isolate X7564_8.85

Isolate X7564_8.85 was sampled from Wishaw in 2010. The ML tree sub-clade shows isolates sampled from many different hospitals, yet it could be posited that this particular isolate originated from Glasgow Royal Infirmary (Figure 4.7a). Isolate X7564_8.85 has 34 SNPs, of which 7 are LSSs. These 7 LSSs are from various different locations; Truro, Belfast, Altnaegelvin, Papworth, Leicester, Sheffield and Glasgow Royal Infirmary. This wide variety of LSS locations obscures any obvious origin hospital. SnAPO provides DOVs for each sampled hospital (Figure 4.7b). This example isolate presents a confusing signal, with no single obvious origin hospital, although the highest DOV (20.7%) is for the same location as that predicted by

the ML tree. Furthermore, the larger DOVs are not clustered together, implying that there might not be one RC that could be the origin. The RC origin DOVs are graphically represented as a geographic heat-map of the UK and Republic of Ireland (Figure 4.7c). Therefore, in this example isolate there is no clear signal for any particular hospital, nor any particular RC. It must be noted that the two RCs with the greatest DOVs are Scotland and Northern Ireland, indicating that the origin of this isolate could be in the north of the UK. However, with the current sampling it is not possible to identify the origin any further. Therefore, in this particular isolate SnAPO has not been able to elucidate much more information than that garnered using a tree based approach.

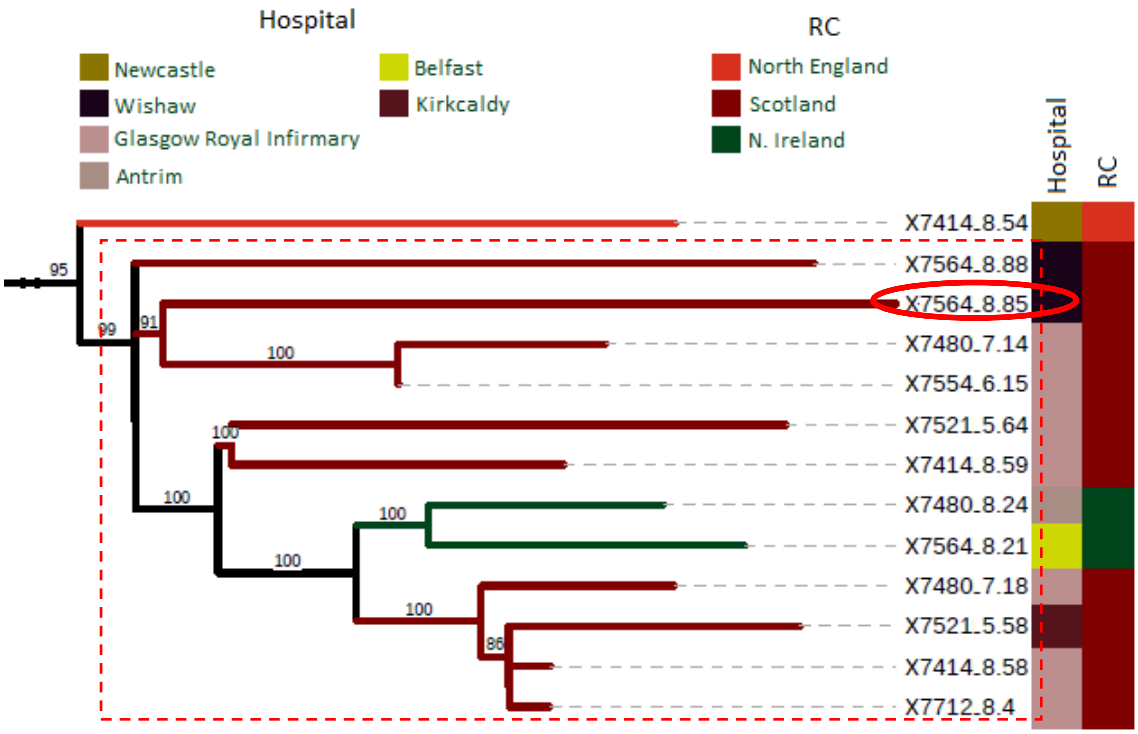


Figure 4.7a. Part of the ML phylogenetic tree depicting the sub-clade (red dashed box) containing isolate X7564_8.85 (red circle). Branch length is proportional to difference in the number of SNPs in the isolates. The numbers above the branches leading to bifurcations are the bootstrap values. If these numbers are absent then this split has a bootstrap value less than 80. The left colour column indicates the hospital the isolate was sampled from, while the right colour column indicates the RC the isolate was sampled from. Branches are also coloured by RC. With this sub-clade clustering it can be difficult to determine the possible origin, but it could be argued that Glasgow Royal Infirmary is the origin hospital.

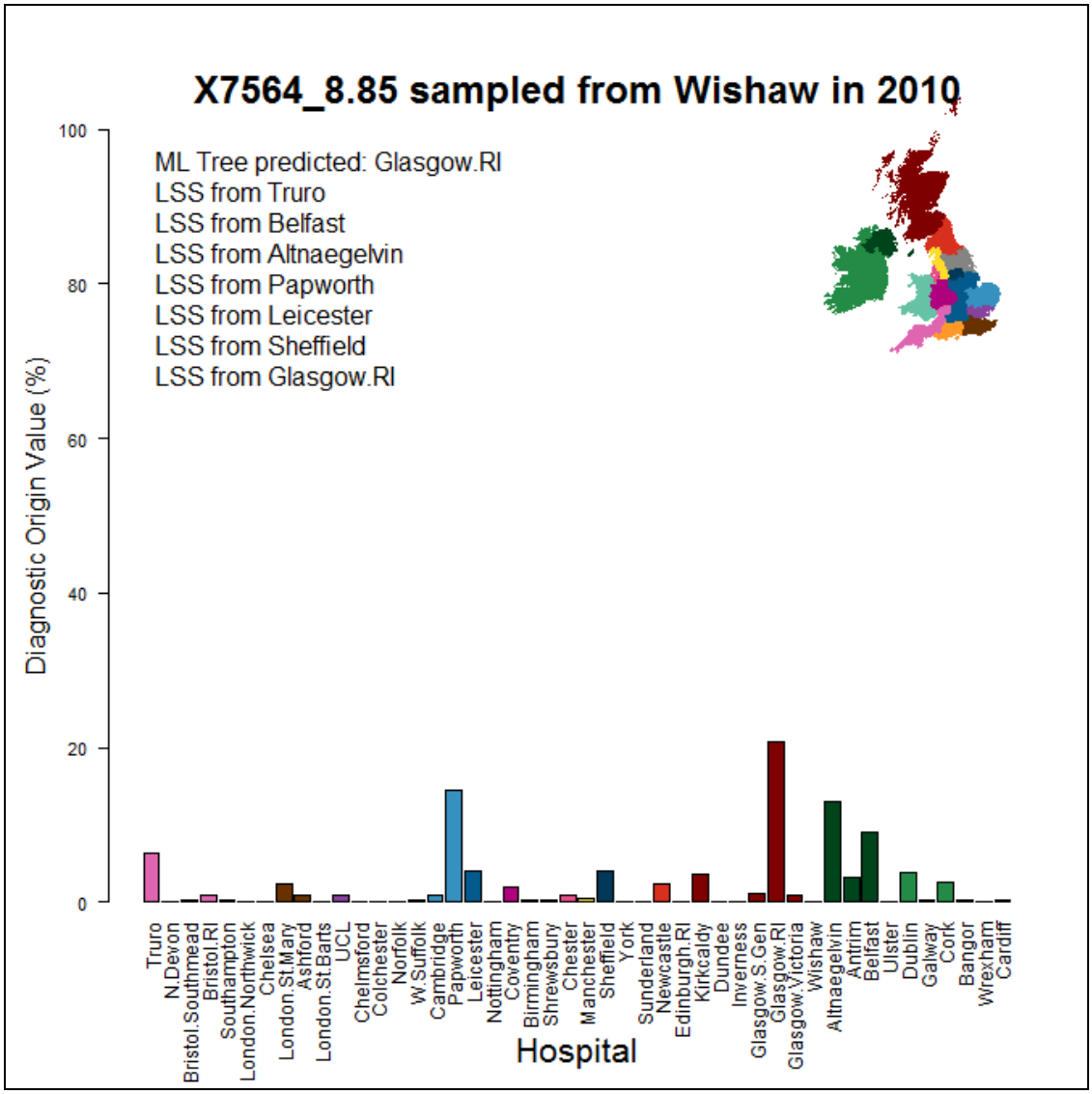


Figure 4.7b. The output for isolate X7564_8.85 shows a signal for Glasgow Royal Infirmary as the origin hospital. However this signal is not clear and the other large DOVs are in geographically disparate locations. The hospitals along the x-axis are laid out in a geographic order, with each Referral Cluster's hospitals adjacent to each other. The y-axis denotes the DOV as a percentage. The bars are coloured by RC, which is provided as a graphical legend.

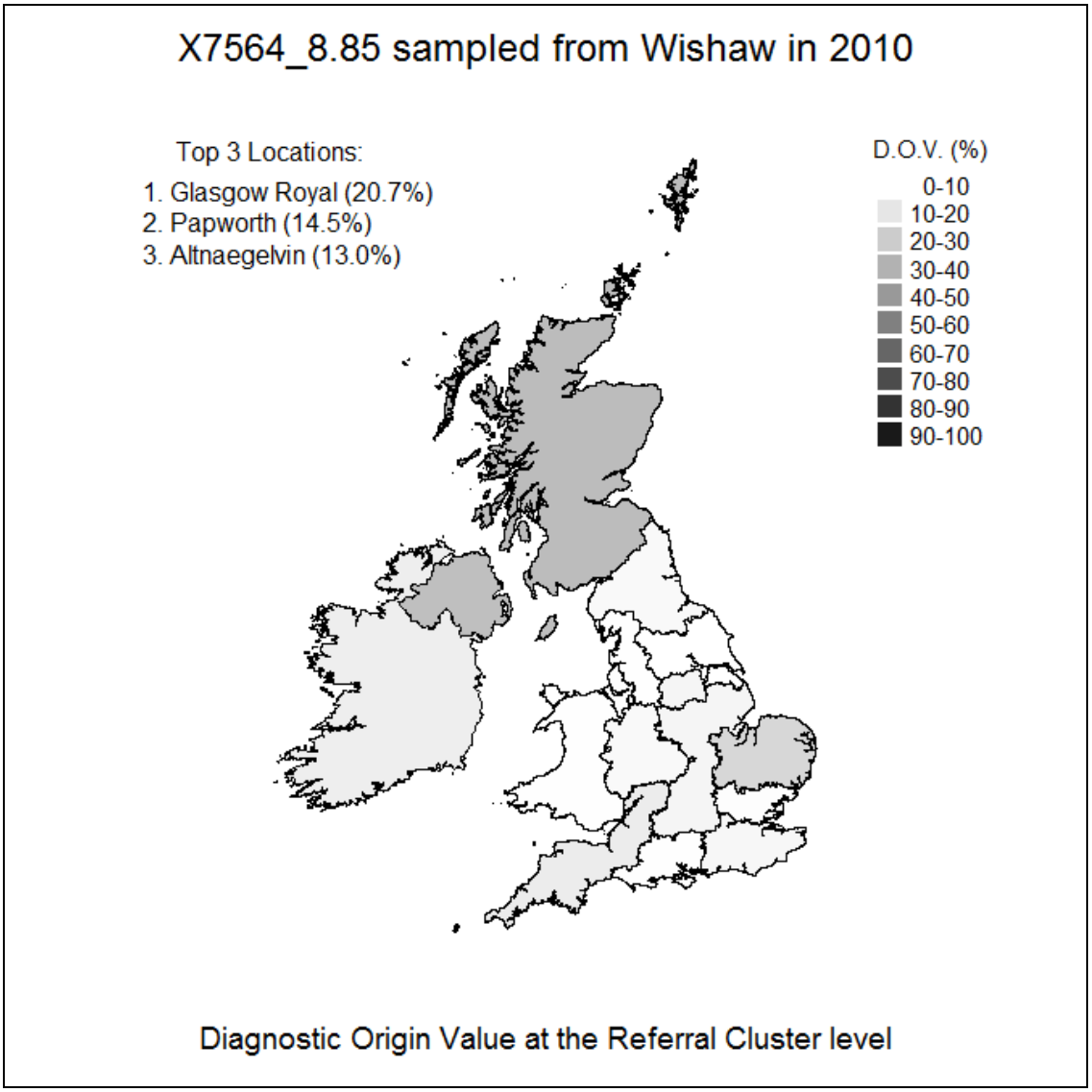


Figure 4.7c. The geographic heat-map of origin for isolate X7564_8.85. There are two RCs that could be the origin of this isolate: Scotland and Northern Ireland. However there is further confusion with the East of England RC showing some indication of origin. Also provided are the top three potential origin hospitals. An increase in opacity indicates a greater DOV for that RC; a key is provided with 10% increments.

4.3.2 Processing the candidate introductions with SnAPO

In Chapter 3 there were 127 possible CIs identified at the hospital geographic resolution. Of these 127 CIs, SnAPO predicted the same posited origin hospital as TAPO in 86 of the isolates (67.7%). The remaining 41 isolates have conflicting predictions. The next step was to see if any of these 41 isolates could have had an incorrect origin assignment from TAPO. Detailed examination of the relevant ML sub-clade indicated that the TAPO prediction was incorrect. There are a variety of reasons, listed as follows and in Appendix C (Supplementary Table C1).

The prevailing reason for the incorrect origin assignment from TAPO was due to the phylogenetic tree construction which groups those isolates that are most similar, regardless of temporal order of sampling. Therefore, if the potential CI was sampled before all, or most, of the other isolates in their sub-clade, then this would preclude that isolate from being a CI. For example, in Figure 4.8 the red isolate could be considered to be a CI from the blue location, due to the phylogenetic clustering. However, if the date at which the isolates were sampled is also known, then the red isolate can no longer be a CI, since it was sampled prior to all the blue isolates. This temporal discrepancy occurred in 22 of the 41 conflicting CIs.

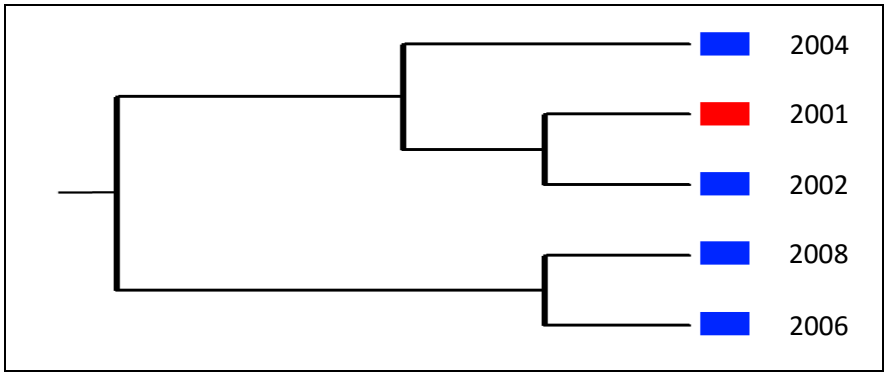


Figure 4.8. A hypothetical sub-clade clustering of five isolates indicating the sampling geographic location (either red or blue) and the year in which the sample was obtained. If the geographic sampling location of these isolates was the only trait considered then it might be posited that the red isolate is an introduction event from the blue location. However, if the date of sampling is then taken into account the red isolate can no longer be considered as a CI since it was sampled prior to all the other isolates.

There were 19 cases in the 41 isolates where the discrepancy between the SnAPO and TAPO origin could not be attributed to the temporal element. However, there are other possible reasons to reject the TAPO-predicted geographic origin. In some instances multiple isolates appeared to be separate introductions on the ML tree but were actually one introduction with subsequent propagation. In a few cases it was the assignation of the sub-clade by the investigator which was erroneous and the sub-clade should have been grouped differently. This final point is due to the subjective nature of the sub-clade groupings.

Although SnAPO can be applied to all isolates, the results will vary (Table 4.1). For some isolates a high DOV is obtained for a single location, while for others there may be multiple locations which could be deemed the origin. A noisy signal will likely remain an issue until sampling and sequencing of MRSA isolates in all hospitals and other healthcare locations becomes commonplace.

Table 4.1. The 127 potential CIs identified by the ML tree were processed using SnAPO. The highest DOV indicates the possible origin location for that isolate. Although each isolate does give an output with a highest DOV, some isolates have a clearer signal. This table summarises the number of isolates which show each maximum value (i.e. the posited origin location of the isolate).

Highest DOV (%)	Number of isolates
0 – 10	0
10 – 20	1
20 – 30	6
30 – 40	10
40 – 50	19
50 – 60	15
60 – 70	13
70 – 80	14
80 – 90	8
90 – 100	0

Using the TAPO method, determining the origin of a potential CI could take some time since it has to be done manually by an investigator with an understanding of phylogenetic properties. Yet, in the novel method developed here, an isolate can be processed in a matter of seconds and provide easily understandable output of a quantifiable nature.

Using these 127 CIs I have shown that SnAPO can predict the origin hospital in the majority of cases. Furthermore, in the minority where there is contradiction, I have shown that the TAPO method could be considered incorrect. Therefore, SnAPO can be used to determine the origin of isolates which are suspected to have a strong geographic signal, based on their ML tree sub-clade clustering.

4.3.3 Comparing the predictions of the test isolates from 2010

The 90 isolates sampled in 2010 were processed using SnAPO and then three independent investigators were asked to determine the possible geographic origin on the ML tree. The investigators were also asked to indicate the level of confidence they have with their decision. The posited geographic origin locations, and the confidence of the decision, of the 90 test isolates determined by all three investigators (using their preferred version of TAPO) and SnAPO were then compared against each other at both the hospital (Appendix C, Supplementary Table C2) and the RC (Appendix C, Supplementary Table C3) geographic resolution.

The posited origin of TAPO from all investigators and SnAPO was compared at the hospital level of geographic resolution (Figure 4.9). It was found that 45 of the 90 isolates (50%) were unanimously predicted to be from the same origin hospital by all investigators and SnAPO. There was still variation in the confidence each investigator assigned the predicted hospital origin, with 28 of the 45 isolates unanimously assigned high confidence and 6 unanimously assigned low confidence by all investigators. The remaining 11 isolates showed variation in confidence by the three investigators. 27 of the 90 isolates (30%) were predicted to have the same origin hospital as SnAPO by at least one investigator. In 20 of the 27 isolates at least one of the investigators could not assign an origin hospital. Where an origin was able to be posited by all investigators, a low confidence was unanimously assigned to 3 of these 27 isolates' origins. In 18 of the 90 test isolates (20%) all investigators predicted a different origin hospital than SnAPO, with 6 of these isolates unanimously assigned a contradicting hospital origin than SnAPO. Some isolates were unable to be assigned an origin hospital by at least one investigator. Those that were assigned an origin hospital were given low confidence estimates. Furthermore, 3 of these 18 isolates could not be assigned to any origin hospital by any investigator. Although one of these 3 isolates has a very low SnAPO DOV maximum (DOV = 23.6%), the other two had DOVs over 50% (DOV = 53.8% and DOV = 53.6%). Finally, ignoring if the posited origin by the investigators contradicted SnAPO it was found that 36 of the 90 test isolates (40%) could not be unanimously assigned to an origin hospital by all three investigators.

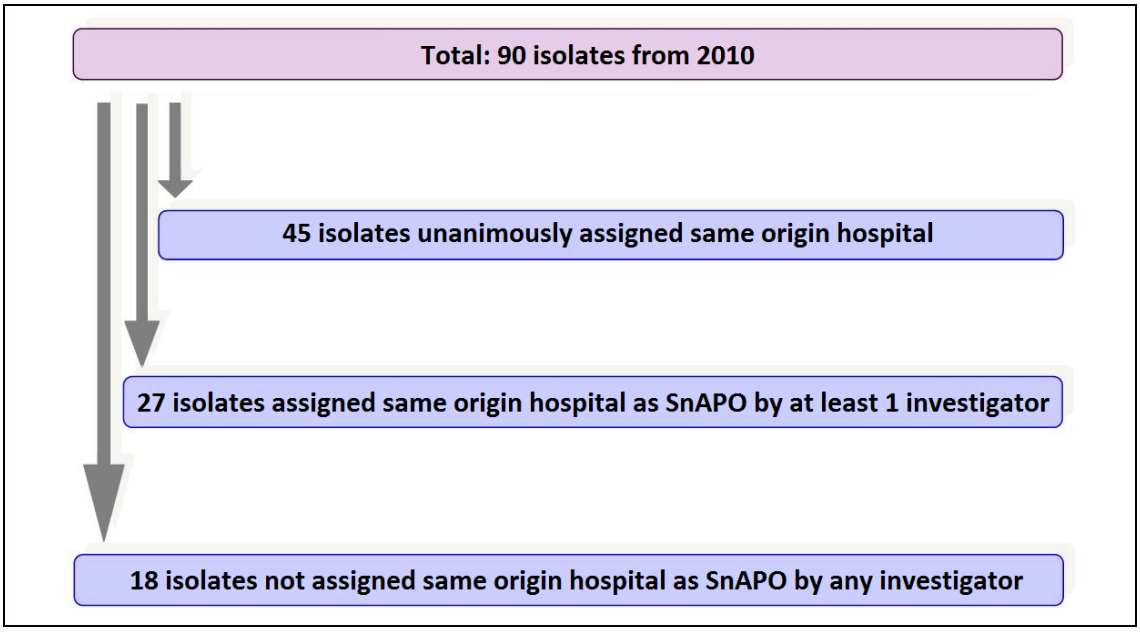


Figure 4.9. The summary of the hospital level comparison between the TAPO method and SnAPO, for the 90 isolates sampled in 2010. Half of the isolates were consistently predicted to have the same origin hospital by all investigators and SnAPO, with a further 27 isolates predicted the same origin hospital as SnAPO by at least one investigator. 18 isolates were not assigned the same origin hospital as SnAPO by any of the investigators. However, there were 6 isolates which were unanimously assigned an origin hospital by the three investigators that contradicted the origin hospital posited by SnAPO.

The posited origin of TAPO from all investigators and SnAPO was then compared at the RC level of geographic resolution (Figure 4.10). It was found that 58 of the 90 isolates (64.4%) were unanimously predicted to be from the same origin RC by all investigators and SnAPO. There was still variation in the confidence each investigator assigned the predicted RC origin, with 44 of the 58 isolates unanimously assigned high confidence and 2 unanimously assigned low confidence by all investigators. The remaining 12 isolates showed variation in confidence by the three investigators. Of the 32 isolates not unanimously predicted an origin, 23 (25.6% of the 90 test isolates) were predicted to have the same origin RC as SnAPO by at least one investigator. In 15 of the 23 isolates at least one of the investigators could not assign an origin RC. Where an origin was able to be posited by all investigators, a low confidence was unanimously assigned to 5 of these 23 isolates' origins. In 9 of the 90 test isolates (10%) all investigators predicted a different origin RC than SnAPO, but no isolate was unanimously assigned a contradicting RC origin than SnAPO. Some isolates were unable to be assigned an origin RC by at least one investigator. Those that were assigned an origin hospital were given low confidence estimates. Furthermore, 3 of these 9 isolates could not be assigned to any origin RC by any investigator. Although one of these 3 isolates has a very low SnAPO DOV maximum (DOV = 25.9%), the other two had DOVs over 50% (DOV = 53.8% and DOV = 53.6%).

These 3 isolates were the same as in the hospital-level investigation. Finally, ignoring if the posited origin by the investigators contradicted SnAPO it was found that 29 of the 90 test isolates (32.2%) could not be unanimously assigned to an origin hospital by all three investigators.

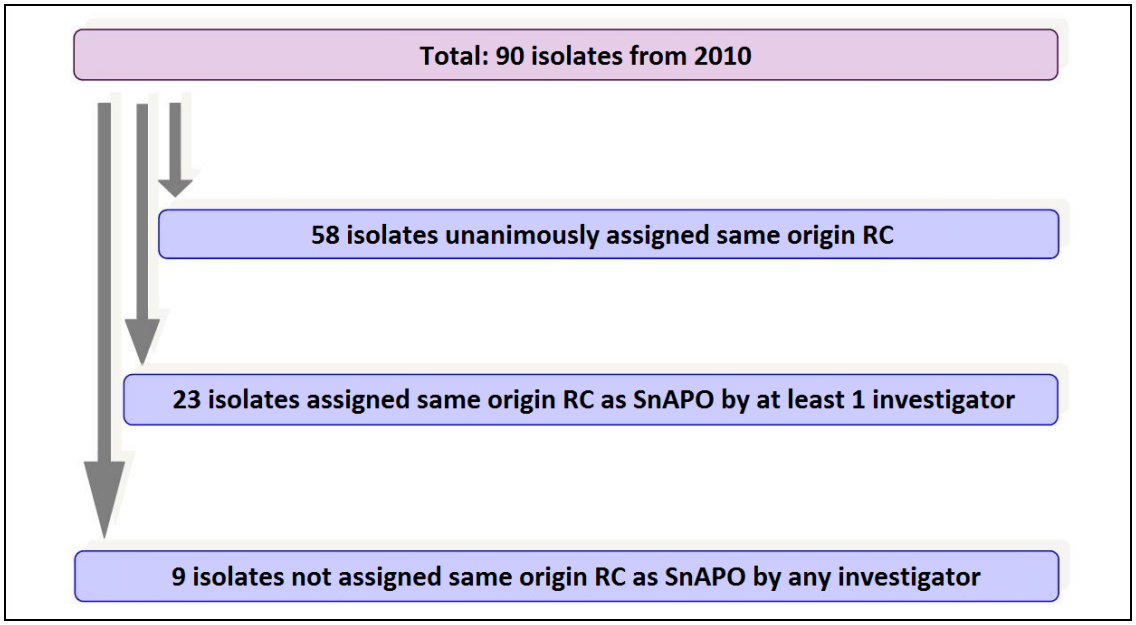


Figure 4.10. The summary of the Referral Cluster (RC) level comparison between the TAPO method and SnAPO, for the 90 isolates sampled in 2010. 58 isolates are consistently predicted to have the same origin RC by all investigators and SnAPO, with a further 23 isolates predicted the same origin RC as SnAPO by at least one investigator. 9 isolates were not assigned the same origin RC as SnAPO by any of the investigators. However, there was no isolate which was unanimously assigned an origin RC by the three investigators that contradicted the origin RC posited by SnAPO.

4.3.4 Determining the speed of SnAPO

The computational speed required for each additional 2010 Test Subset isolate processed using SnAPO was determined under two conditions; when using the full Comparison Subset and half the Comparison Subset. It was observed that doubling the number of the isolates in the dataset does not significantly increase the time required to process each isolate, with 90 isolates processed using SnAPO in 38.37 seconds and 42.04 seconds, for a dataset size of 466 and 932 isolates respectively (Figure 4.11). The time reported here does not include the time required to create the database, nor the time required to transform the data into a useable format for SnAPO.

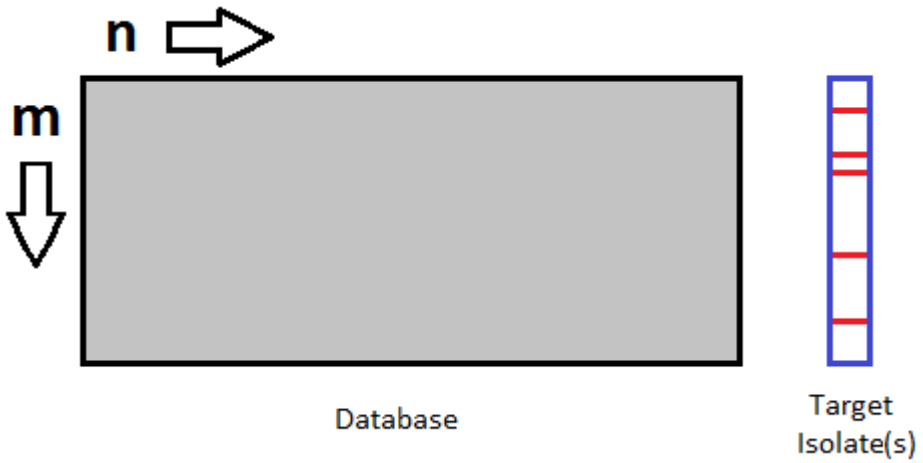


Figure 4.11. A database with *n* isolates and *m* SNP positions was used to process the target isolates through SnAPO. 90 isolates, sampled in 2010, were used to test the effect of doubling *n*, while keeping *m* constant at 5461 SNP positions. With *n* = 466 SnAPO took 38.87 seconds while with *n* = 932 SnAPO took 42.04 seconds to process the 90 test isolates. This shows that for this database, doubling the number of isolates in the database does not double the time required by SnAPO. The time determined here does not include any time required to transform the data into the correct format for SnAPO. Furthermore, each target isolate only expresses certain SNPs (*s*), denoted in this diagram by red lines, and this may further help speed up the process since SnAPO will only need to examine the previous incidence of those SNPs.

4.3.5 Processing the 2011 and 2012 isolates

The 17 isolates from 2011 and 2012 were processed with SnAPO with no knowledge of the sampling location. The SnAPO predicted origin location was compared with the actual sampling location (Table 4.2). The Diagnostic Origin Values (DOVs) for the prediction origin location varied greatly between the target isolates (from 13.1% to 73.0%), with 12 isolates showing a maximum DOV less than 40%. Therefore, for these 12 isolates the origin location cannot be confidently concluded through SnAPO. The large majority (n = 14) of SnAPO predicted origin locations do not match with the sampling location, however this is likely due to the large number of isolates from previously un-sampled locations.

Table 4.2. The isolates taken from 2011 and 2012 were processed by SnAPO with the sampling location metadata removed. The predicted SnAPO origin location was then compared to the sampling location. A number of the sampling locations in 2011 and 2012 were not previously included in the database. Although the isolates were processed in the order that they were sampled, the table is arranged by decreasing maximum DOV.

Isolate Code	Date of Isolation	SnAPO Hospital Prediction	Max Hospital DOV (%)	SnAPO RC Prediction	Sampling Hospital	Sampling RC
X8113_5.88	02/10/2011	Cambridge	73	RC8	Cambridge	RC8
X8140_1.86	16/02/2012	Cambridge	60.6	RC8	Peterborough	RC8
X7915_6.3	08/02/2011	Papworth	51.9	RC8	Yarmouth	RC8
X8113_5.91	23/02/2012	Cambridge	47.2	RC8	Cambridge	RC8
X7915_6.20	26/12/2011	Papworth	45.1	RC8	Ipswich	RC8
X8728_5.50	12/10/2011	Cardiff	37.7	RC16	Southend	RC4
X7915_6.25	25/01/2012	Cambridge	35	RC8	Peterborough	RC8
X7915_6.19	06/10/2011	Cambridge	33.6	RC8	Watford	RC4
X7915_6.1	26/08/2011	Papworth	33	RC8	Yarmouth	RC8
X7915_6.5	25/02/2011	Cardiff	32.7	RC16	Basildon	RC4
X8113_5.90	18/01/2012	Cambridge	31.5	RC8	Cambridge	RC8
X8113_5.89	11/01/2012	Leicester	31.4	RC9	Cambridge	RC8
X8113_5.87	26/05/2011	London St. Mary	26.9	RC1	Cambridge	RC8
X7915_6.18	14/09/2011	London St. Mary	23.4	RC1	Watford	RC4
X7915_6.6	04/09/2011	Cork	18.8	RC13	Basildon	RC4
X8113_5.92	16/03/2012	Manchester	18.3	RC12	Cambridge	RC8
X7915_6.15	27/03/2011	London St. Mary	13.1	RC1	Watford	RC4

Of the 5 isolates which show maximum DOVs higher than 40%, 2 of them (X8113_5.88 and X7915_6.25) were sampled from Cambridge and predicted to come from Cambridge. The location predicted by SnAPO for an isolate from a previously un-sampled hospital would likely be ambiguous. However there are 2 cases where this is not the case; isolates X7915_6.3 and X7915_6.20 (sampled from Yarmouth and Ipswich respectively) both have high maximum DOVs and are predicted to originate from a hospital within the same referral cluster as the sampling hospital. However, these isolates represent the first samples of their respective hospitals in the dataset and therefore the SnAPO-predicted location would always be different from the sampling location. The last isolate with a maximum DOV higher than 40% (X8140_1.86) is sampled from Peterborough yet SnAPO predicts it to come from Cambridge with a DOV of 60.6%.

4.3.6 Processing isolates from Holden *et al* (2013)

The 10 isolates sampled in 1991 all show ambiguous SnAPO output, with multiple low peaks and none over 40%. This is as expected for the isolates from 1991, since the SNPs they express are probably in many different locations in the isolates in the Bi-allelic Dataset.

There are a total of 56 isolates sampled from non-UK countries in 2006 and 2007. Of these 28 isolates have high maximum DOVs for locations other than their sampling location, and so might be considered to be introduction events. However, it is probable that these are not true introduction events, since they are sampled from locations which have not been sampled before. The more likely scenario is that the SnAPO posited origin location is the one which contains the highest number of similar isolates.

There are a total of 41 isolates sampled from within the UK in 2006 and 2007. Of these, 16 isolates have high maximum DOVs for locations other than their sampling location (Table 4.2), and could be considered to be introduction events. However, the majority of these isolates were from locations within the UK that not been sampled previously, and so one could not confidently assert that these isolates are representative of true introduction events. Only one isolate (X07_1361_K sampled in Dundee in 2007) was from a location that had been previously sampled and showed a high maximum DOV for a location that was not its sampling location. Therefore, this could be a true introduction event.

Table 4.3. There are 16 isolates from the Holden *et al* (2013) study sampled in the UK which show high DOVs (more than 40%). However, only one of these isolates (X07_1361_K) was sampled in a location that had been previously sampled in the Bi-allelic Dataset. Therefore, this is the only isolate which could be considered to be an introduction event, since all the other isolates came from previously un-sampled locations. Further sampling in those locations, coupled with an expanding database would likely reduce this issue.

Isolate	Year	Country	Sampled	SnAPO Origin	Max DOV (%)
X07_1361_K	2006	UK	Dundee	Kirkcaldy	53.6
X07_2384_Y	2007	UK	Stornoway	Manchester	88.4
X07_5739_N	2007	UK	Glasgow	Manchester	80.3
X6401_6_11	2007	UK	London	London St. Mary	60.3
X6401_6_13	2007	UK	London	Papworth	54.0
X6401_6_14	2007	UK	London	West Suffolk	41.4
X6401_6_16	2007	UK	London	Papworth	61.0
X6401_6_17	2007	UK	London	Papworth	58.8
X6401_6_18	2007	UK	London	London St. Mary	49.1

X6401_6_19	2007	UK	London	Papworth	61.0
X6401_6_6	2007	UK	London	London St. Mary	61.4
X6401_6_8	2007	UK	London	Papworth	56.6
X6401_6_9	2007	UK	London	Ashford	58.4
X6401_7_18	2007	UK	Manchester	Sheffield	40.7
X6401_7_20	2007	UK	Buckinghamshire	Papworth	49.3
X6401_7_21	2007	UK	Lincolnshire	Cambridge	40.1

A single location was then chosen, and after each isolate from that location was processed with SnAPO it was added to the dataset. The 14 isolates sampled from London in 2007 were chosen. Of these 14 isolates there are 7 isolates with a maximum DOV higher than 40% (Table 4.3), with 3 isolates showing their origin to be within London; either London St. Mary or the new London location. Furthermore, those isolates sampled later show an increasing DOV for the new London location.

Table 4.3. The 14 isolates from the Holden *et al* (2013) study sampled in London were re-analysed with SnAPO, with each isolate being added to the dataset once it was processed. Therefore, in this situation there is slightly more confidence that the isolates with high DOVs could actually have originated from the location posited by SnAPO. There were 7 isolates which showed high DOVs over 40%, and of these there are 4 isolates which show a possible origin outside of London, indicating possible transmission events. The remaining 3 isolates with high DOVs show origins within London, which could indicate transmission events within London or, due to large geographic region encompassed by London, this could be the specific sampling location. Furthermore, there is a slight increase towards the later isolates for an increasing DOV for the London location.

Isolate	Year	Sampled	SnAPO Origin	Max DOV (%)	DOV for London (%)
X6401_6_15	2007	London	Papworth	25.6	0
X6401_6_13	2007	London	Papworth	53.3	1.419994251
X6401_6_16	2007	London	Papworth	56.6	6.147975477
X6401_6_18	2007	London	London St. Mary	49.0	0.184030105
X6401_6_10	2007	London	Papworth	29.4	2.416891968
X6401_6_7	2007	London	London St. Mary	26.8	5.358041074
X6401_6_6	2007	London	London St. Mary	57.5	6.681392931
X6401_6_17	2007	London	Papworth	37.1	34.00228515
X6401_6_12	2007	London	Chelmsford	26.1	0
X6401_6_14	2007	London	West Suffolk	41.4	0
X6401_6_9	2007	London	Ashford	58.4	0.587900542
X6401_6_8	2007	London	London	36.9	36.87311994
X6401_6_11	2007	London	London	37.0	37.0379504
X6401_6_19	2007	London	London	41.7	41.68147137

The majority of the London isolates did not show much change between the first version of SnAPO, which did not add the isolate to the dataset, and the second version, which did add the isolate to the dataset (Figure 4.12). Some isolates showed a slight decrease in maximum DOV in the second version, and there are three isolates (X6401_6_13, X6401_6_17 and X6401_6_11) which showed a decrease in maximum DOVs from above 40% to below 40% between the two versions of the SnAPO process. This is likely due to the inclusion of an extra category for London, which may now contain some of the DOV signal.

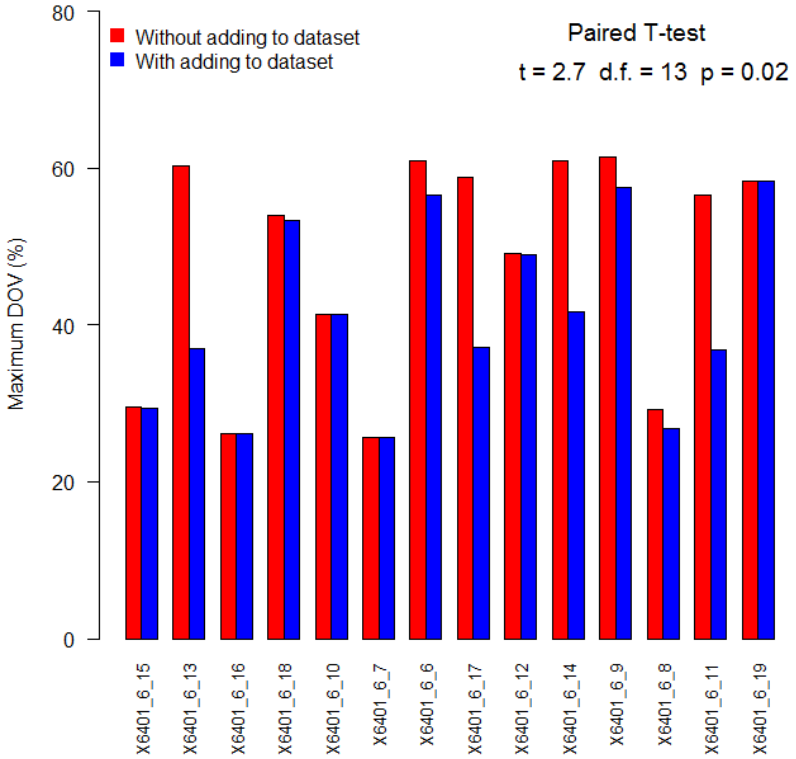


Figure 4.12. The 14 isolates from Holden *et al* (2013) sampled from London in 2007, were processed twice with SnAPO. In the first version each isolate was not added to the dataset, while the second version each isolate was added to the database after it was processed. Although the majority of the 14 isolates retained the same maximum DOV, there are a few which have a decreased maximum DOV in the second version. This is likely due to the presence of an extra category for London. A paired t-test shows that there is a significant difference in the DOVs of each version, though this is likely driven by a few isolates with large differences (such as, X6401_6_13)

4.3.7 Including SnAPO in an analysis pipeline

It was stated previously (Section 4.3.1) that processing an isolate using the traditional tree-based approach is a lengthy process, requiring a person with an understanding of phylogenetic techniques and tree construction programs. Furthermore, once the tree is built,

which in itself could take some time depending on the phylogenetic method employed, the investigator could spend considerable time classifying each sub-clade and isolate in the tree, especially in trees of the size of the current and future MRSA datasets and collections. If WGS becomes routine in healthcare institutions the ever expanding database would soon make this approach impractical.

The novel method presented in this paper, SnAPO, provides a quick way to circumvent this issue and, if added to an existing sequencing pipeline, would be cheap to implement. Furthermore, the output can be easily displayed to facilitate interpretation, with a minimal of technical knowledge required to understand the data. As has been mentioned, the characteristic examined in this study was the geographic location, but SnAPO could be applied to any genetic characteristic where the appropriate metadata are known; for example, particular virulence phenotypes. With the drive towards including greater active surveillance and WGS in healthcare institutions it is possible that including SnAPO as part of the analysis package would greatly facilitate the interpretation and identification of outbreaks within a shorter time scale. The use of SnAPO is currently limited to those isolates which have already been identified to a particular clonal complex or sequence type using traditional methods, such as MLST or PFGE. Therefore, a possible analysis pipeline could look like that described in Figure 4.13.

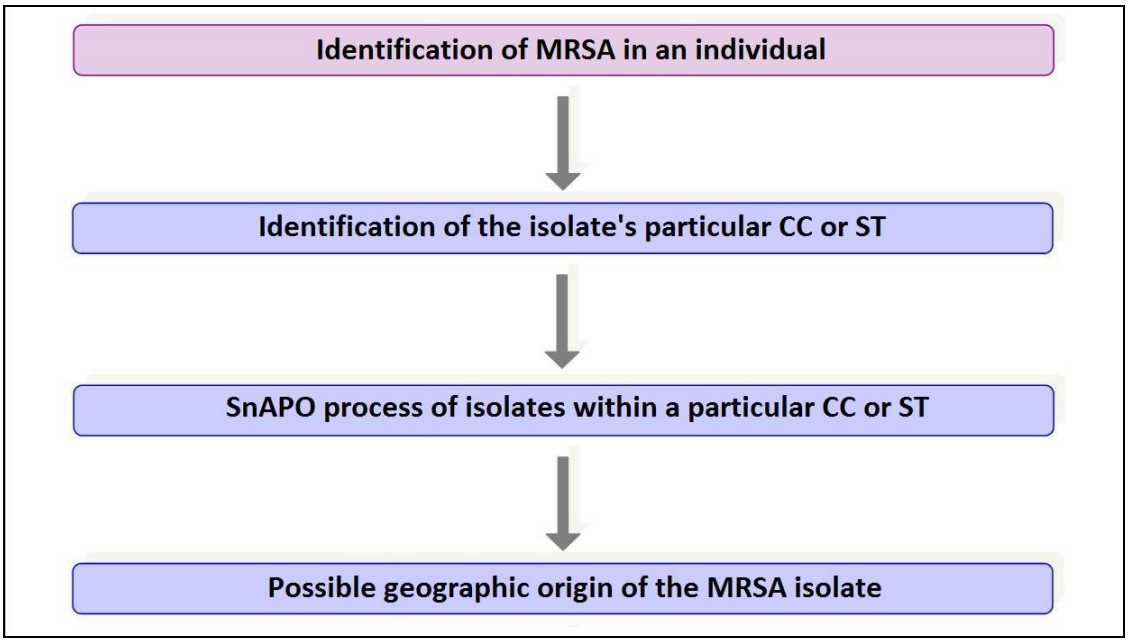


Figure 4.13. A theoretical analysis pipeline for the identification of an MRSA isolate sampled from a healthcare facility or laboratory. The identification of the particular Clonal Complex (CC) or Sequence Type (ST) would be done with conventional methods, such as MLST or PFGE. Once segregated into different lineages, the isolates can be processed using SnAPO to identify possible geographic origin of an isolate, within the confines of the sampling.

The final output of SnAPO could be represented graphically on one sheet for each isolate. The following page (Figure 4.14) shows an example output that might be produced by a healthcare institution using this method. The sampling effort in each hospital has been included as a final panel in the output. A selected number of extra examples are included in Appendix E.

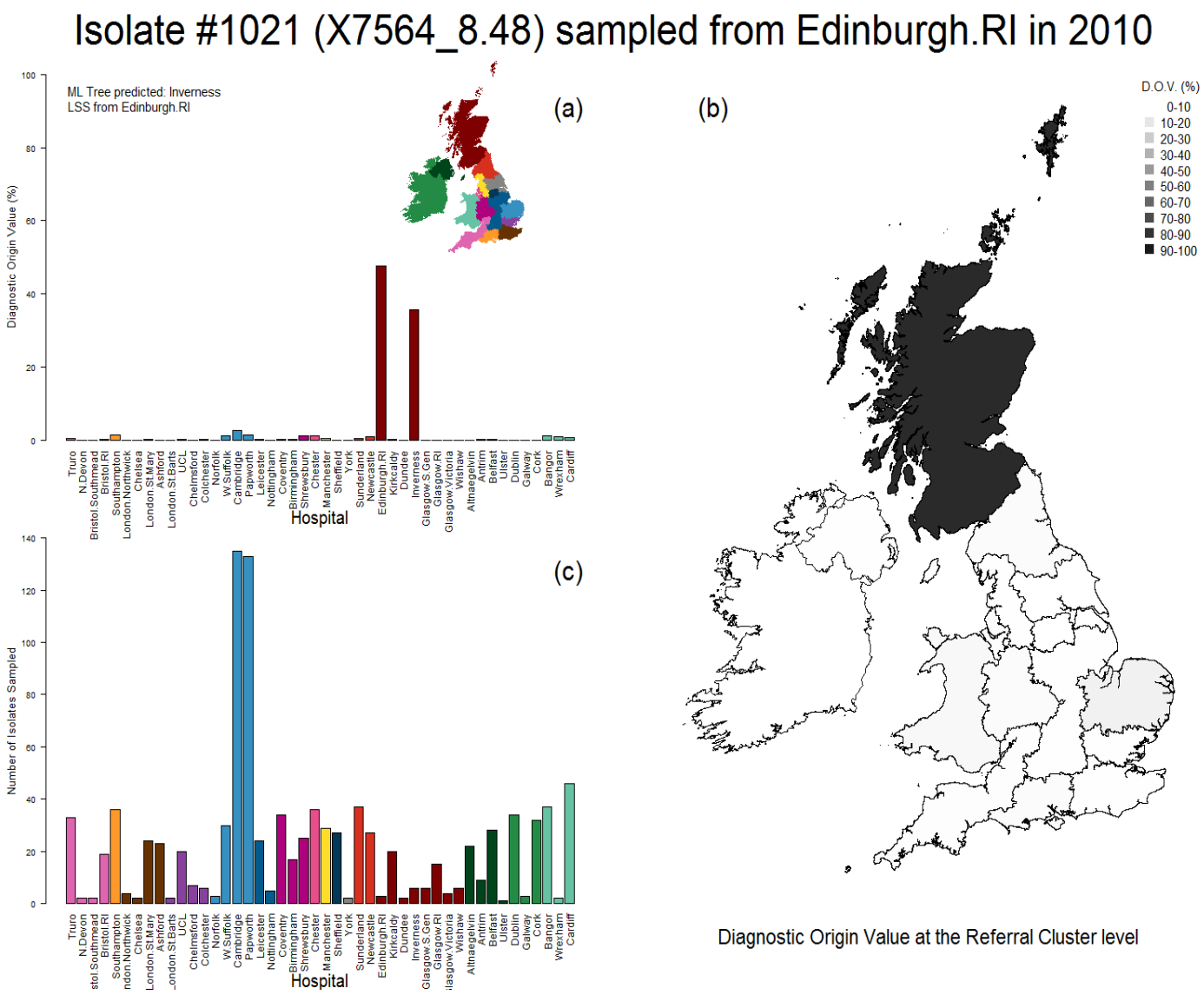


Figure 4.14. An example output that might be produced by a healthcare institution using SnAPO with a collated online database of MRSA genomes. The first panel (a) shows the SnAPO output as a DOV percentage. In this example (isolate number 1021 in the Bi-allelic Dataset) the isolate appears to have been originated from Edinburgh Royal Infirmary or Inverness. Any LSSs the isolate expresses are displayed, and the origin location as predicted by TAPO. A coarser scale of geographic resolution is displayed (b) using Referral Clusters (RCs). Finally, the sampling effort in each hospital is included as a final panel (c) in the output. The bars in (a) and (c) are coloured by the RCs (provided as a graphical legend in a) and are ordered by their geographic proximity. The shading in (b) is split into 10% bins. The key is provided in the top right corner.

4.4 Discussion

I have presented a novel method for determining the origin of isolates of MRSA based on the SNPs that an isolate harbours. I have termed this method the SNP-based Assignment of Pathogen Origin (SnAPO). I have shown that SnAPO is fast and easily interpretable. It is also objective, in that for a given focal isolate and a given dataset SnAPO will always return the same output, unlike in phylogenetic analyses. However the interpretation of the output for isolates with no clear DOV may remain a subjective issue. Furthermore, I have shown that any isolate can be processed using SnAPO, not just those that are potential Candidate Introduction events. The entire process could easily be automated and would replace the labour intensive phylogenetic approach. It is important to note that SnAPO does not try to predict future transmission events, but rather provides a rapid summary of where it might have originated within the confines of the geographic sampling.

SnAPO is computationally fast with only a slight increase in time required when the dataset size is doubled for this database. For a dataset of this size, SnAPO can process a target isolate within a second. However, this may change if a centralised repository of MRSA genomes and the implementation of SnAPO in healthcare institutions is developed. Therefore, the relationship with time for SnAPO could be considered to be at least O(nm). It could be argued that since SnAPO only considers the SNPs expressed (*s*) in a given target isolate then the time relationship could be O(ns), and since s < m this would be faster than O(nm).

Traditional phylogenetic approaches show a larger increase in computational time required when doubling the dataset size. For example, NJ is $O(n^3)$ (Studier & Kepler, 1988) and the construction of the NJ phylogenetic tree (Section 2.4.1, Figure 2.3) took approximately 15 minutes, though programs such as RapidNJ may decrease the time order of certain steps to $O(n^2)$ (Simonsen *et al*, 2011). Alternatively, ML is NP-hard and therefore would take even longer than NJ to compute for large datasets (Chor & Tuller, 2005), and the construction of the ML tree (Section 2.4.1, Figure 2.4) took approximately two weeks to complete. Therefore, if SnAPO is O(nm) then this is faster than ML and could be comparable to NJ, while if SnAPO is O(ns) then this would be faster than both phylogenetic methods.

The time consideration for SnAPO obtained here does not include the time required to process the data into a format compatible with the SnAPO process. This formatting would slightly increase the computational time required. Conversely, the time it would be required to manually assign an origin location to each isolate once the phylogeny has been created is not included which, as discussed, is a time consuming process. Furthermore, increasing the

number of SNPs (*m*) would also increase the time required, regardless if using SnAPO or a phylogenetic approach. However, I believe that SnAPO would still be computationally faster, even if these additional processes were included into the time required to run SnAPO versus traditional phylogenetics. Additionally, it might be possible that the dataset used was not sufficiently large enough to determine the true time order of SnAPO and so further work in this direction with larger datasets would be informative. Conversely, most of the mathematical and computational methods used in epidemiology are often very complex and require trained personnel to implement and interpret the results. Therefore, there might be a bottleneck in the analysis pipeline caused by the limited number of trained personnel available for interpretation of other methods. Implementation of easily understandable methods, such as SnAPO, that do not require extensive training would reduce the bottleneck and help combat the spread of a pathogen.

The comparison between three investigators' separate TAPO interpretations and SnAPO of the 2010 isolates indicates that TAPO might be too conflicting in its subjectivity. This subjectivity would be due to each investigator's interpretation of the sub-clade clustering and phylogenetic metadata. Furthermore, it is possible that SnAPO can provide higher confidence to origin locations than that available to phylogenetic methods. Although not all MRSA isolates will be able to have an exact origin assignation using SnAPO, I advocate its use for the majority of isolates. Furthermore, SnAPO can process every isolate in near real time and could identify transmission events that might be missed on the phylogenetic tree.

However, the majority of the 17 isolates sampled in 2011 and 2012 have DOVs less than 40%. This is likely due to the fact that only Cambridge had been sampled prior to 2011 and so it is not unreasonable to have ambiguous SnAPO output for newly sampled hospitals. An interesting case is isolate X8113_5.92 which was sampled in Cambridge but has a low DOV positing Manchester as the origin. Therefore, it is possible that this may be an indirect introduction event from Manchester to Cambridge. It might be possible to resolve this issue with more comprehensive sampling.

There are 5 isolates with DOVs higher than 40% in the 2011 and 2012 isolates. Of these, 3 could be potential candidate introductions at the hospital geographic resolution, since they are posited to come from a different hospital than their sampling location with a DOV higher than 40%. This approximately matches what was found when using the 91 isolates in the 2010 Test Subset, where 24 of them (26.3%) were possible hospital resolution candidate introduction events. However, of these 3 possible candidate introductions, 2 were the first

samples from their respective hospitals and therefore the sampling hospitals was not a possible outcome. Therefore, SnAPO had to posit a different origin hospital to the sampling hospital, and so this is not necessarily indicative of a transmission event. Although possible that these are true transmission events, it might be not be prudent to conclude that these are actual transmission events. The last isolate with a DOV higher than 40% was sampled in Peterborough but was predicted to come from Cambridge. This is the second isolate from Peterborough in the dataset, and so one can be slightly more confident that this is a true transmission event. However, further sampling at that location will be required.

Future work could look into improving upon the arbitrary 40% DOV threshold which may help identify an isolate from an unallocated location, or alternatively to avoid using SnAPO in a given location until a sufficient minimum number of isolates have been sampled from that location. Finally, it is worth noting that SnAPO does place all the isolates with DOVs higher than 40% into the same RC as where they were sampled from, indicating that it is doing a reasonable job with the information available.

Additionally, I have shown that it is possible to process isolates from a different dataset using SnAPO. Some of the isolates in Holden *et al* (2013) were processed with SnAPO and it was found that there were a number of isolates with high maximum DOVs. Some of these isolates were from non-UK countries, and therefore one could not confidently say that they were introduction events from the SnAPO posited UK hospital to that country. However, there are a couple of interesting situations. Firstly, there are 13 isolates from Germany that have high DOVs. Of these 13 isolates, 7 of them are predicted to have arisen from Manchester with high DOVs and the rest are from Papworth. Therefore, although it is unlikely this is representing direct introduction events from either Manchester or Papworth to Germany, it could be that these UK locations and the German ones historically shared the same isolates. In a similar situation, there are 6 isolates from Singapore with high DOVs for UK locations. All of these isolates are predicted to come from London St. Mary, and all of them have very similar DOVs (within a percentage). Therefore, this could represent some historic connection between the isolates in Singapore and those in London St. Mary.

Many of the UK locations in Holden *et al* (2013) were locations that do not appear in the Bi-allelic Dataset. Therefore, unless there are multiple samples from one location it is not possible to be confident in the output. The majority of isolates in Holden *et al* (2013) were from locations not found elsewhere in the Bi-allelic Dataset. However, a number of the isolates which were sampled from previously un-sampled locations and have high DOVs, are posited to

have an origin outside of their sampling Referral Cluster. Therefore, this lends further support to the possibility that these might be actual introduction events. However, as previously stated, since these isolates each represent the first incidence of that location in the dataset it would be unwise to conclude that they are representative of introduction events.

The isolates sampled from London in 2007 which were processed twice with SnAPO, once with each target isolate not added to the dataset after processing, while the second time with each isolate being added to the dataset and used to compare to the following isolates. Although some of the maximum DOVs might be lower, the overall output could be considered to be more informative since the addition of each target isolate to the dataset enables more confidence to be placed in the outputs. However, it should be noted that the new London location is still not very specific and the MRSA isolates could have originated from differing locations within London.

Therefore, further sampling from these locations, coupled with a growing database to which these isolates can be added, would provide an informative comparison database and increase the clarity of the SnAPO output. I believe that the creation of this online database would be an important next step in making SnAPO a viable practical tool. Although much more work needs to be done to achieve a practical application of SnAPO, it is promising that viable, if not yet completely trusted, output can be obtained from isolates sampled in different locations for a different study.

A crucial part of SnAPO is the development of an online database of MRSA that each individual healthcare institution can upload sequences to as they are processed. Although screening for MRSA is becoming more commonplace in hospitals in the UK (Dancer, 2008), sequencing of those isolates have still to be implemented in the majority of hospitals. However, each additional isolate added to the database would increase the resolution of the output, allowing for a "snowball effect" to occur. With this in mind, I stress that this must be the next step in order to unlock the full potential of SnAPO. The application of SnAPO coupled with a collated online MRSA database, would enable the rapid determination of the geographic origin, and hence the spread, of MRSA isolates. The containment and eradication of an outbreak of a virulent MRSA strain is an expensive process (Kanerva *et al*., 2007) and so identification of the outbreak early will allow better focusing of limited resources and personnel. Although the geographic location has been used in this paper, this method could be applied to non-geographic characteristic. For example, if certain SNPs are always in isolates

with a particular virulence phenotype then this characteristic could be used to determine the course of treatment a particular isolate may be susceptible to.

Unfortunately, there remain limitations to SnAPO. Currently the sampling effort is not extensive enough to allow for complete accuracy, but this could be rectified by more universal sampling and sequencing of MRSA. Therefore, the geographic origin of each isolate is limited to those locations sampled. Other than the sampling issue, the main problems arise with homoplasy, recombination, convergent evolution and compensatory mutations. These processes could lead to SNPs arising in disparate isolates that have had no recent connection, which could lead to an erroneous conclusion of the geographic origin. I have attempted to mollify any effect of this by ensuring that any one SNP cannot skew the result too heavily, but it is not inconceivable that multiple SNPs in disparate isolates have arisen independently. If such a situation was to arise it would be difficult to determine the origin location. It is likely this will present as an isolate with a particularly noisy origin signal.

Although homoplasy and recombination may result in a noisy origin signal, with conflicting output and no location as the immediate obvious origin, this noisy signal could also be due to the isolate originating from outside the country or from locations within the UK and Ireland that have not been sampled. A noisy output can be attributable to the isolate only containing SNPs which are fairly common in all locations, i.e. old SNPs. Therefore, it is possible that an isolate with only old SNPs could be from a country outside of the UK and Ireland. In Chapter 5 I will investigate the application of a Bayesian approach, and in Chapter 6 I will explore the use of other genetic information such as indels.

SnAPO is, to the best of my knowledge, the first implementation of a method which examines the SNP incidence to automatically determine the geographic origin of an isolate without using a phylogenetic tree. There have been other attempts to automate this process, with a first generation analysis program developed by Brossette *et al*. (2000) called the Data Mining Surveillance System (DMSS). This system was not focused on MRSA, but rather was attempting to be applicable to any nosocomial infection. However, this system still required further interpretation of the analysis, and the authors note that it was only the first step in the process to automate surveillance systems. Alternatively, Mellmann *et al*. (2006) have developed a similar process using *spa* type data; attempting to develop a fast diagnostic of MRSA epidemic outbreaks. Although they have reported some success, the advent of cheap whole genome sequencing (WGS) has opened many other avenues of investigation. Using an extensive number of SNPs, as is now possible, gives much greater flexibility. This was seen in a

number of recent studies which used WGS to elucidate genetic information from a group of MRSA isolates (e.g. Harris *et al*., 2010, 2013; Price *et al*., 2014; Tong *et al*., 2015). These studies all used phylogenetic analysis techniques to determine possible transmission routes and geographic origins of the MRSA isolates. Therefore, it would be interesting to see if the application of SnAPO to the datasets used in these studies would contradict the phylogenetic findings.

However, the SnAPO method would be used once the isolates have been identified to a particular clonal complex or sequence type by traditional methods, such as MLST, PFGE or *spa* typing. Therefore, the incorporation of conventional methods, WGS and SnAPO into one analysis process would be the optimum solution. With that in mind, I also recommend that further exploration of other potential genetic signatures, such as indels, should be carried out in order to possibly obtain an even clearer signal of an isolate's origin.

4.5 Conclusion

The development of a novel method, SnAPO, has been shown to successfully identify the possible geographic origin of an MRSA isolate within the confines of the sampling. Using SnAPO connected with an ever expanding online database of MRSA isolates from around the country, it could be possible to confidently, and rapidly, determine the isolate's geographic origin. However, if the target isolate is the first sample obtained from a location then SnAPO would posit an origin location that would be different from the sampling location. In these situations, it might be erroneous to conclude that that isolate is indicative of a transmission event even if there is a high maximum DOV.

Furthermore, SnAPO is more objective than a phylogenetic approach, in that the output remains the same for a given focal isolate and a given dataset. The limitations of SnAPO arise when processing isolates with no clear origination signal; these outputs require more subjective interpretation. A threshold DOV value could be instigated, over which one might be more prepared to trust SnAPO (e.g. origin predictions with DOV values higher than 40%). I will investigate whether these limitations may be overcome by alternative methods, such as Bayesian analysis in Chapter 5, or by using alternative genetic characteristics, such as indels in Chapter 6.

The work in this chapter shows that signature genetic signals can be teased out of the genome and has laid the foundation for future work in this direction. If the sequencing of all MRSA isolates in healthcare institutions becomes commonplace, due to the ever decreasing cost of WGS, then the implementation of SnAPO as part of the analysis pipeline would provide a simple and easily interpretable graphical output of the isolate's geographic origin. This information would be informative for determining spread of particular MRSA strains and applying the limited resources available.

5 A Bayesian approach to SnAPO

5.1 Introduction

In the previous chapter I introduced a novel method (SnAPO) which was able to determine where a focal isolate may have originated from, within the confines of the Bi-allelic Dataset, for the majority of the test isolates. In this chapter, I move away from the heuristic SnAPO method and explore a SNP-based Bayesian classification approach to determine the geographic origin of an isolate. The Bayesian inference approach is an established non-heuristic statistical method of classification. Since SnAPO is a heuristic method, it would be prudent to investigate if a Bayesian approach concurs with SnAPO. There has been some success previously using Bayesian approaches, either at identifying transmission events or clustering similar isolates together. An attempt to determine transmission events using genomic data and Bayesian inference was conducted by Didelot *et al*. (2014). This study was successful in determining genetic diversity, but there was still considerable uncertainty when identifying transmission events. The authors conclude that Bayesian reconstruction may be a useful starting point, but traditional epidemiology would remain the main process of identifying transmission events. However, the approach used by Didelot *et al*. (2014) was computationally intensive. Furthermore, with an expanding dataset traditional Bayesian clustering methods, such as those implemented by STRUCTURE or BAPS software, might be impractical (Jombart *et al*., 2010). Therefore, a less computationally demanding approach would be required for practical implementation in a healthcare institution. In this chapter this will be attempted to be achieved by considering each SNP individually. It is possible that using an individual SNP approach will allow the identification of the geographic origin of an isolate using realistic processing requirements.

In summary, in this chapter a SNP-based Bayesian inference approach is constructed to determine the geographic origin of an isolate. This will move away from the heuristic SnAPO method described in Chapter 4 and follow established statistical practices. The identification of the SNPs used in this thesis is described in Section 2.2

5.2 Dataset and methods

In this chapter it is assumed, for convenience, that the posited geographic origin for each of the 90 2010 Test Subset isolates obtained in Chapter 4 is the correct geographic origin. Therefore, in the methods described in this chapter the results obtained here will be compared with those obtained in Chapter 4. To simplify the narrative the results obtained in Chapter 4 will be termed the "Primary SnAPO result".

5.2.1 Bayesian classification for determining an isolate's geographic origin

The method presented here was developed by Jukka Corander and Richard James, and implemented on the Bi-allelic Dataset by James Sciberras. It is a SNP-based Bayesian classification approach for determining the geographic origin of an isolate. In this situation there are a finite number of hospitals (k = 46), each of which is considered a class into which the data can be assigned. Unlike the heuristic SnAPO method described in Chapter 4, in this chapter an established Bayesian classification process will be used (Corander *et al.*, 2011).

We will let h denote hospital and s denote SNP locus. In the Bi-allelic Dataset each focal isolate comprises of a vector $\boldsymbol{x}$ of 5469 bi-allelic SNP loci s with $x_s \in \{0, 1\}$, where 0 corresponds to the nucleotide which is a SNP mutation and 1 corresponds to the non-SNP nucleotide. Our goal is to determine the probability that $\boldsymbol{x}$ comes from hospital h. To determine this we will require a prior distribution and a likelihood function. The prior distribution $p(h)$ would come from the number of isolates in the dataset sampled from each hospital and the likelihood function will be obtained from the dataset. The formulation of the likelihood function is explained in greater detail further on in this section. We use the dataset to generate the likelihoods π_{hs}^{Z} that at locus s in hospital h we will find nucleotide Z (0, 1). For example, at locus s = 1 in a hypothetical focal isolate we have x_1 = 0. From the dataset we find that the likelihood that there is a 0 at locus 1 in hospital h = 10 is 0.03, while in h = 11 it is 0.6. Therefore, locus 1 is implying that hospital 11 is the more likely origin. However, we need to combine all 5469 loci s in focal isolate $\boldsymbol{x}$. We do this by multiplying the likelihoods from all the loci, using the π_{hs}^{Z} values for the nucleotide Z that appears at each locus in our focal isolate. We do this for each hospital in turn. Once we normalise over all possible classes, which in this situation are the 46 hospitals, we will have the posterior probability distribution over h for the origin of $\boldsymbol{x}$.

We derive π_{hs}^{Z} values from the dataset as follows. We first take the expected (μ) occurrence of Z at locus s in hospital h as the measure of likelihood. If there are n_h isolates

from hospital *h* in the dataset, and n_{hs}^{Z} of them have nucleotide *Z* at locus *s*, a simple non-Bayesian estimate of probability would be $\mu_{hs}^{Z} = n_{hs}^{Z}/n_h$. However, we can improve on this estimate by implementing a prior belief in the rarity of the SNPs, using the conjugate prior distribution. In this case the appropriate conjugate prior is the Beta distribution. The shape of the Beta distribution is governed by two hyperparameters (a, b); the distribution mean is $a/(a+b)$. The method for determining the values of *a* and *b* in this situation are provided later on in this section. The posterior distributions is also a Beta, with parameters $n_{hs}^{0} + a$ and $n_h - n_{hs}^{0} + b$ and mean (i.e. likelihood)

$$\pi_{hs}^{0} = \frac{n_{hs}^{0}+a}{n_h+a+b}. \qquad \text{Equation 5.1}$$

Since $n_{hs}^{0} + n_{hs}^{1} = n_h$, then

$$\pi_{hs}^{1} = 1 - \pi_{hs}^{0} = \frac{n_{hs}^{1}+b}{n_h+a+b}. \qquad \text{Equation 5.2}$$

We now have a prior, which represents a hypothetical extra $(a+b)$ isolates in each hospital; *a* of which are SNPs, *b* are non-SNPs at each locus *s*. All loci are assumed to be of equal importance, so we combine the evidence from all 5469 bi-allelic SNP loci by calculating the products of the various π_{hs}^{0} or π_{hs}^{1}, guided by the presence or absence of a SNP at each locus in the focal isolate vector ***x***. This is repeated for each hospital in the dataset. This gives a likelihood function for the vector ***x***. We now include the prior probability *p(h)*, determined by the number of isolates at hospital *h* divided by the number of isolates over all hospitals, to give the Bayesian Origin Value (BOV) for hospital *h* (Equation 5.3):

$$BOV_h = p(h)\prod_{s=1}^{S}\pi_{hs}^{Z}. \qquad \text{Equation 5.3}$$

We normalise BOV_h by the sum of all possible classes. For 46 hospitals this is trivial compared to other Bayesian classification applications which might have very large number of classes. We now have the posterior distribution over *h* for the focal isolate ***x***. We will term this the Bayesian Diagnostic Origin Value (BDOV). It should be noted that the BDOV for a focal isolate usually shows one hospital with a much larger BDOV than any other hospital. Therefore, examining the log BDOVs would reveal more detail of the BDOVs for the other hospitals.

Thus far we have not yet defined the hyperparameters *a* and *b*. We know that SNPs are very rare and so a neutral choice (e.g. *a* = *b* = 1) would not be appropriate. The prior

distribution is normally constructed through knowledge obtained from sources other than the system in question. However, in this situation we will use the dataset to inform us of the prior distribution. Although this is counter-intuitive, it is a valid approach since it can be a useful approximation (*Jukka Corander personal communication*, 2015) and there is precedent (for example, see the analysis of cancer death rates in the US in 1980-1989 in Gelman *et al*., 2014). The distribution of SNP rarity of all loci in the database (see Section 2.4.2) has a very small mean (0.0065) and mode (0.0021), with the majority of SNPs only present in two isolates. This is consistent with the proportion of total SNPs in the whole Bi-allelic Dataset (0.0066). With two known values (mean and mode) and two unknown values (*a* and *b*) we can derive *a* = 1 and *b* = 228 to obtain a Beta distribution which is consistent with the distribution of the SNP rarity (Figure 5.1). This means that with this prior expectation we would find 1 in 229 isolates harbouring a SNP at any chosen locus.

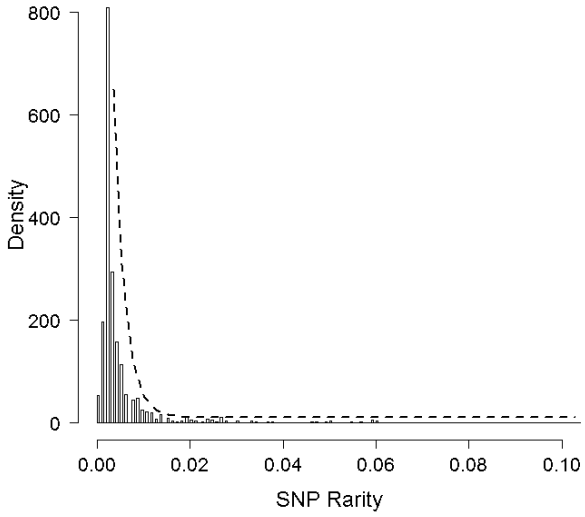


Figure 5.1. The rarity of the distribution of the SNPs in the dataset can be used to inform us of the prior distribution. The mean (0.0065) and mode (0.0021) of this distribution can be used to obtain the two hyperparameters (*a* and *b*). In this situation this gives us values of 1 and 228 for *a* and *b*, respectively (black dashed line). SNP rarity is determined by the number of isolates harbouring a SNP over all isolates in the dataset.

Finally, as with SnAPO, each time we analyse an isolate we add this isolate to the dataset before moving on to the next isolate. Therefore, the addition of each isolate would subtly change the π_{hs}^{Z} for a SNP locus *s* at a hospital *h*. Furthermore, the addition of each isolate will change the *p(h)* for the hospital where it was sampled.

To test the Bayesian classification approach the Primary SnAPO results of the 90 2010 Test Subset isolates were compared with the results from the method described here. For each test isolate I determined which hospital shows the highest value in both methods. I determined the number of test isolates which showed the same posited origin hospital in both methods, and the number of test isolates which showed the SnAPO posited origin hospital as one of the top three Bayesian posited origins. Furthermore, the output of those that do show the same posited origin hospital in both methods was examined to determine the level of

agreement by comparing the value obtained through each method. For example, even if both methods posit the same origin hospital there may be a large discrepancy between the BDOV and the SnAPO DOV. It must be noted that this would be subject to interpretation by the investigator, since the SnAPO method is heuristic and therefore it is not yet clear what constitutes a “high” DOV. Nevertheless, very low DOVs would still imply a noisy signal. Finally, those test isolates which did not show the same posited origin hospital were examined and attempted to identify the cause, where possible.

5.3 Results

5.3.1 Bayesian classification for determining an isolate's geographic origin

It was found that 72 of the 90 isolates (80.0%) sampled in 2010 showed the same posited hospital origin with the Bayesian approach as with SnAPO. The majority of the test isolates show a single large BDOV, while SnAPO is much more varied (Table 5.1). Therefore, agreement in both methods does not automatically indicate that the origin posited by either is the true origin of the isolate. The noisiness of the SnAPO signal must be taken into consideration, leading to some subjective interpretation by the investigator on the validity of the results of each method. The BDOVs and the SnAPO DOVs for all isolates is supplied in Appendix D (Supplementary Table D1).

Table 5.1. The value for the posited origin for the 90 isolates in the 2010 Test Subset using both the SnAPO (DOV) and the Bayesian inference approach (BDOV) was determined. Most of the test isolates exhibit high values when using the Bayesian approach, but there is variability in the value when using SnAPO. This indicates that agreement of the two methods might not be sufficient for a confident conclusion that the posited origin is the true origin. Interpretation of each output by the investigator would be necessary. The BDOV and DOV values can be represented either as proportions or percentages

BDOV or DOV of posited origin (%)	SnAPO isolates	Bayesian isolates
0 – 10	0	0
10 – 20	5	0
20 – 30	19	0
30 – 40	10	0
40 – 50	16	0
50 – 60	28	2
60 – 70	6	2
70 – 80	4	2
80 – 90	2	2
90 – 100	0	82

Those cases that did not show the same posited origin in both methods (Table 5.2) were examined and found that for the majority of the test isolates the discrepancy arose due to the presence of Location Specific SNPs (LSSs) for the posited origin hospital. SnAPO allows the LSSs and rare SNPs more influence in determining possible origin, therefore if an isolate harbours an LSS this would increase the final DOV for that hospital. However, the Bayesian approach has no such weighting, and this appears to be the reason for the different posited origins between the two methods in many of the test isolates. Furthermore, it appears that

those isolates which have an unclear output in SnAPO often do not show the same posited origin in both methods. Finally, the Bayesian posited hospital origin was one of the top three posited hospital origins by SnAPO in 87 of the 90 2010 Test Subset isolates (96.7%). This indicates that it is rare for both methods to completely disagree in the posited origin.

Table 5.2. The 18 isolates of the 2010 Test Subset were examined which showed a different posited origin when using the Bayesian approach compared to SnAPO. The differences can be mainly attributed to the presence of LSSs or the very unclear SnAPO output obtained.

Strain Code	Sampling Hospital	SnAPO Posited Origin	Bayes Posited Origin	Description of difference
X7564_8.19	Belfast	Antrim	Belfast	The SnAPO DOV for Belfast is very similar to Antrim, indicating it could be from either.
X7748_6.69	Antrim	Altnaegelvin	Sunderland	The output from SnAPO is very unclear and not possible to confidently assign origin hospital based on SnAPO.
X7748_6.70	Antrim	Antrim	Sunderland	SnAPO output shows very low DOV for Sunderland. An LSS for Antrim may be causing difference observed.
X7564_8.22	Belfast	Belfast	Dublin	Dublin appears as the second highest DOV in SnAPO, with similar value to Belfast.
X7564_8.31	Cardiff	Cambridge	Cardiff	Cardiff is second highest DOV in SnAPO, while the isolate harbours an LSS for Cambridge.
X7564_8.75	Sunderland	Newcastle	Sunderland	SnAPO output is very ambiguous, with Sunderland appearing as second highest value.
X7083_1.29	Cambridge	Cambridge	Papworth	Papworth appears as second highest in SnAPO output, and the isolate harbours an LSS for Cambridge.
X7564_8.68	Glasgow Southern General	UCL	London St. Mary	London St. Mary appears as the second highest DOV in SnAPO, with a value close to that of UCL. The isolate also harbours an LSS for UCL.
X7748_6.66	York	Cambridge	Sheffield	Sheffield appears as second highest DOV in SnAPO, with a value close to that of Cambridge. The isolate also harbours an LSS for Cambridge.
X7564_8.42	Cork	Dublin	Cork	The DOV for Cork is very similar to that of Dublin in the SnAPO output.

X7564_8.89	Wishaw	Wishaw	Manchester	DOV for Manchester is second highest in SnAPO. Isolate also harbours an LSS for Wishaw.
X7083_1.30	Cambridge	Papworth	London St. Mary	London St. Mary is second highest DOV in SnAPO. Isolate also harbours an LSS for Papworth.
X8728_5.39	Edinburgh Royal Infirmary	Kirkcaldy	Belfast	SnAPO output is very unclear and no hospital can be confidently identified as the origin. Belfast does appear as second highest DOV in SnAPO.
X8728_5.38	Edinburgh Royal Infirmary	Kirkcaldy	Belfast	SnAPO output is very unclear and no hospital can be confidently identified as the origin. Belfast does appear as third highest DOV in SnAPO.
X7915_8.13	Colchester	Norfolk	Cardiff	Isolate harbours only 8 SNPs, one of which is an LSS for Norfolk.
X7915_6.14	Colchester	Papworth	Cardiff	SnAPO output is unclear and no hospital can be confidently identified as origin. Isolate also only harbours 11 SNPs.
X7564_8.48	Edinburgh Royal Infirmary	Edinburgh Royal Infirmary	Inverness	Inverness has second highest DOV in SnAPO, close in value to Edinburgh Royal Infirmary. Isolate also harbours LSS for Edinburgh Royal Infirmary.

An example output of the type a healthcare institution may create if they were to employ both the SnAPO and the Bayesian method is provided (Figure 5.5). A select few examples of the test isolate outputs can be found in Appendix F. Implementation of both methods would allow greater flexibility in the interpretation of the output, especially in situations where the two methods do not concur. Furthermore, for those isolates which do concur then the inclusion of the SnAPO output would indicate whether the isolate shows a noisy signal, which would not be visible when just using a Bayesian approach.

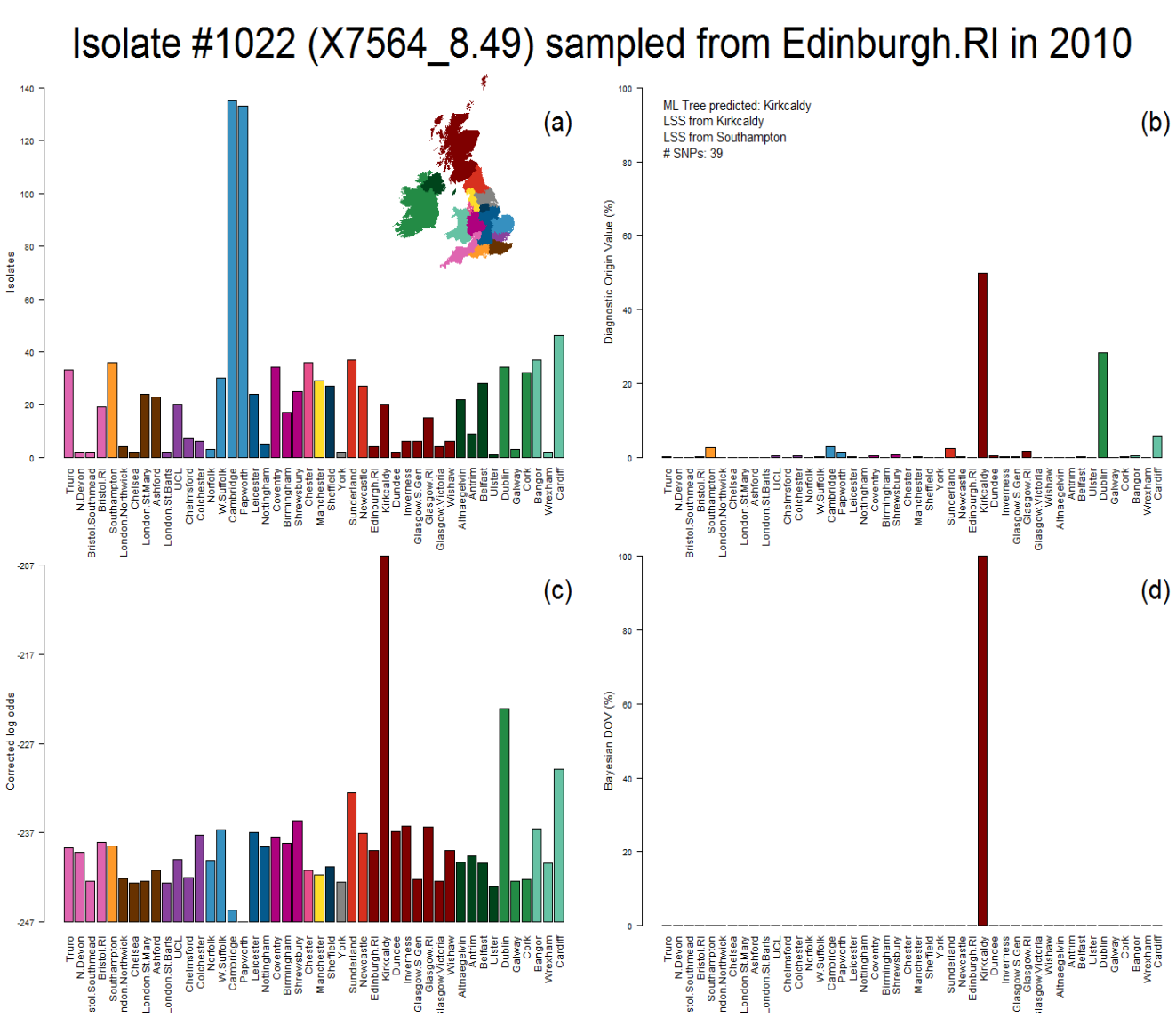


Figure 5.5. An example of the possible output that could be created by a healthcare institution to determine the origin of an MRSA isolate. The sampling effort in each hospital is shown (a). This is what was used to obtain *p(h)* in Equation 5.3. Both the SnAPO (b) and the Bayesian method (d) are included, allowing for informed interpretation should the two methods disagree on the posited origin. The log version of the Bayesian output is also included (c), since more variation is visible in this graph. Finally, this example was chosen to illustrate how SnAPO (b) often has multiple hospitals with high DOVs, while the Bayesian approach usually shows only one hospital with a high value. In the top left corner of (b) the origin predicted by the ML tree, any LSSs harboured by the isolate, and the number of SNPs the isolate harbours are all displayed. In all plots the bars are coloured by Referral Cluster, provided as a graphical legend in (a).

5.4 Discussion and Conclusion

I have explored the use of a Bayesian classification approach to determine the geographic origin of an isolate within the confines of the dataset, and that this approach concurs with the Primary SnAPO result in the majority of the 2010 Test Subset isolates. In the Bayesian classification approach described here it was assumed that each isolate may have originated from within the hospitals sampled; this is termed a "supervised approach" (Corander *et al*., 2013). However, it is likely that some isolates may have originated from un-sampled hospitals or from outside the UK. A semi-supervised approach is possible by introducing an empty extra class with no defined prior likelihood function (Corander *et al*., 2011). This would allow those isolates which are unlikely to arise from any currently defined class to be assigned as originating from this new, extra class. In this situation, this would not force an origin from the hospitals in the dataset on those isolates which may have originated from un-sampled locations or from out of the UK. This semi-supervised approach could likely be the next step in the development of this method.

Bayesian inference is an established statistical principle which judges, in probabilistic terms for each hospital, whether there is data similar to that observed in the focal isolate in the dataset (Corander *et al*., 2011; Gelman *et al*., 2014a). Therefore, it might appear that using a heuristic method such as SnAPO is unnecessary. However, our Bayesian method has its own limitations. Firstly, it assumes that all the SNP loci are independent since the output is achieved from the product of each SNP locus. As mentioned in Section 2.4.3, many of the SNPs are not independent and furthermore, many of the SNPs are nested within each other. SnAPO would be more appropriate if one assumes that the SNPs are non-independent, since the impact of each SNP on the result is weighted by their rarity in the dataset. It is important to mention that not all SNPs are linked, nor are all SNPs nested. Therefore, neither the Bayesian approach described in this chapter, nor the SnAPO method are perfectly appropriate. Secondly, the Bayesian inference approach treats each SNP locus as equally important. Therefore, the information available in the rarer SNPs might be obscured by the large number of non-SNP loci. This is reminiscent of the issues, described in Section 2.5.1, that may preclude F_{ST} from being an appropriate measure of similarity. Finally, our Bayesian classification is more computationally demanding than SnAPO. Currently this is not a serious issue, however with an ever-expanding dataset this might prove to be problematic for practical implementation. For these reasons, and the high level of agreement between the established Bayesian inference and the heuristic SnAPO methods, I suggest that the implementation of the heuristic SnAPO method is a viable alternative. It is likely that the most appropriate method lies somewhere

between these two approaches. Combining these two novel methods would be the next step in their development, either by displaying both outputs or by creating some intermediary version.

6 Expanding SnAPO & exploring limitations

6.1 Introduction

In Chapter 4 I showed that a novel method (SnAPO) was able to posit a geographic origin for any isolate, although some isolates resulted in clear output signals, while others were ambiguous. In Chapter 5 the principles developed for SnAPO were modified for use in a Bayesian approach with similar success. In this chapter I explore and expand the use of SnAPO to determine if some of these ambiguous cases might be better described using alternate genetic information. I also explore some of the possible limitations of SnAPO; namely, the potential lack of robustness to changes in the dataset, and the possible decrease in relevant information available in older isolates for determining the posited geographic origin for an isolate.

This chapter will consist of three topics. Firstly, I wished to see if it was possible to use other genetic information to obtain a predicted geographic origin for an isolate. To do this the use of nucleotide insertions or deletions (indels) was explored. As with SNPs, there is a very slight chance of two indel mutations of exactly same size and nucleotide composition occurring at exactly the same genomic position, indicating that they might be valid markers for identification of relatedness (Väli *et al*., 2008). Furthermore, there appears to be a pattern of covariation between SNPs and indels within and between species (Chen *et al*., 2009). These factors imply that indels could be used in a similar way to how SNPs were used in Chapter 4 to identify a possible geographic origin of an isolate.

Many studies on sequence variation focus on SNPs, recombination and microsatellites (Albers *et al*., 2011), but the development of the use of indels as genetic markers has lagged behind (Väli *et al*., 2008). This lag is likely due to the increased difficulty in identification of indels compared to other genetic markers. There is often a high error rate in mapping indels and occasionally it is not possible to uniquely map an indel to a genomic position; for example, if a repeat unit in a tandem repeating sequence is deleted then it is difficult to determine where in the repeat sequence this deletion occurred (Albers *et al*., 2011). Furthermore, an increase in the genomic indel rate often corresponds to an increase in the number of indels

due to sequencing errors (Albers *et al*., 2011). Regardless of these difficulties there has been a recent increase in the use of indels, as they can be viewed as an untapped resource of genetic information. For example, a method to differentiate eukaryotic species has been successfully developed which relies on the indel variation in rRNA (Pereira *et al*., 2010) indicating that there is sufficient genetic variation in indels to distinguish organisms at the species level. In recent years there has been an increase of the use of indels in studies of humans (Mills *et al*., 2006) and model species such as *Drosophila melanogaster* (Ometto *et al*., 2005).

In general, indels are not as common as SNPs in most species, with indels occurring approximately twenty times less often than SNPs in bacteria (Chen *et al*., 2009). However, in most organisms they also tend to be widely spread across a genome (Väli *et al*., 2008), though there is variation between coding and non-coding regions with more indels occurring in non-coding regions (J. Q. Chen *et al*., 2009). Indels can be identified based on their size with a comparable monetary cost to that of SNP identification (Väli *et al*., 2008). There are two broad methods to identify indels (summarised in Albers *et al*., 2011). The first involves the assembly of short reads *de novo* and the comparison of these reads to a reference sequence. The second method maps each short read independently of other reads to the reference sequence. Variation is identified as a difference between the reads mapped to that genomic location and the reference sequence at that genomic location.

The second topic discussed in this chapter will focus on the effect that changes to the composition of the dataset may have on the posited origin location of an isolate. This topic will consist of two parts. In the first part it is determined whether the sampling effort in the dataset affects the designated origin location of an isolate. As was mentioned in Chapter 3 the Location Specific SNPs (LSSs) and the rare SNPs may only be rare due to the sampling effort in the Bi-allelic Dataset. Therefore, it is possible that alterations to the Bi-allelic Dataset may create different posited origin hospitals for a target isolate. This dependency on the sampling effort may contribute to a lack of robustness to changes in the dataset in the SnAPO method. Therefore, this needs to be explored. The robustness of the method is tested by randomly removing isolates and observing if this alters the predicted geographic origin of any of the 2010 Test Subset isolates. The second part of this topic will focus on the possible influence of the age of the isolates in the dataset. It is possible that some of the data used in this thesis are too old to provide meaningful information in SnAPO. Therefore, there might be some merit to removing old isolates to obtain a clearer signal, if they are not providing any positive contribution to the determination of a geographic origin.

Finally, the last topic examined in this chapter is the possible variation of information signal attributable to individual SNPs, and the possible repercussions if one were to change the SNPs an isolate expresses. Therefore, simulations were run which replaced increasing numbers of SNPs in each target isolate with a randomly selected SNP and investigated the changes, if any, in the SnAPO output. This information was used to determine how many SNPs can be modified before the origin location is obscured. For this test, the isolates sampled in 2010 which showed maximum DOVs higher than 40% were used (see Chapter 4).

In summary, in this chapter I will explore three topics related to SnAPO and how it might possibly be improved and expanded. I will do this by focusing on the finer, and potentially more informative, scale of hospital geographic resolution origin. Firstly, I will attempt to implement SnAPO on a different genetic variation characteristic; indels. Secondly, I will determine if SnAPO is robust to changes in the composition of the dataset, and if the older isolates may impact on the clarity of the final SnAPO output. Finally, I will determine how many SNPs can be lost in an isolate before the origin location is obscured.

6.2 Dataset and methods

As in Chapter 5, in this chapter it is assumed that the posited geographic origin for each of the 90 2010 Test Subset isolates obtained in Chapter 4 is correct. Therefore, in this chapter the results obtained here will be, once again, compared with those obtained in Chapter 4. The results obtained in Chapter 4 are once again called the "Primary SnAPO result". Additionally, the identification of the SNPs used in this thesis is described in Section 2.2.

Furthermore, for each of the isolates processed in Chapter 4 the Diagnostic Origin Value (DOV) for the posited geographic origin and the minimum number of hospitals required to obtain 90% of the whole DOV signal was determined. A low number of hospitals required to obtain 90% of the signal is indicative of a clean signal, usually implying there is one hospital with a high DOV. A large number of hospitals required to obtain 90% of the DOV signal is indicative of a very noisy output. These two metrics (the DOV of the posited geographic origin, and the number of hospitals required for 90% of the DOV signal) can give some indication of the clarity of an output; a higher DOV with a low number of hospitals is indicative of a cleaner signal. It must be noted that the 90% threshold is an arbitrary value. This value was chosen since it was observed that many isolates showed a tiny DOV for every hospital in the dataset and taking the number of hospitals required to obtain 100% of the DOV signal usually resulted in all the hospitals, which was not informative.

6.2.1 Amending SnAPO for an indel dataset

The indels were identified using Dindel (Albers et al., 2011) with reads being realigned around indels using the GATK toolbox and the ratio used was 0.65. Although there is SNP data for 1022 isolates, there is only indel data for 1009 isolates. There are 88 isolates sampled in 2010 with both indel and SNP data. Furthermore, there is only 899 unique indel positions. However, each indel position can have up to four unique indels. An indel may range from a single insertion or deletion, to a large number of nucleotide insertions or deletions. Therefore, a new dataset was constructed with 1009 isolates and 899 unique indel positions. This will be termed the Indel Dataset.

The SnAPO method (Chapter 4) was amended for use on the Indel Dataset; i.e. each indel in a focal isolate was examined and it was determined where those indels were previously found. This information was then used to determine the origin hospital for the 88 test isolates from 2010 for which there is indel data. This indel result was compared to the Primary SnAPO result to determine if the predicted hospital is the same, or if the Primary SnAPO result hospital is one of the top three hospitals posited by the indel data.

The Indel Dataset and Bi-allelic Dataset was then combined for those isolates which have both indel and SNP information to examine if the addition of the indel data gives a clearer prediction of the origin of the isolate. This will be termed the Combined Dataset. The 88 2010 isolates with both indel and SNP data are once again used as the test cases. The Combined Dataset was processed using the SnAPO method and the results were compared to the Primary SnAPO result. I investigated if the posited origin hospital is the same, or if the Primary SnAPO result is one of the top three Combined Dataset results.

6.2.2 Testing the robustness of SnAPO to changes in the dataset

Due to the nature of SnAPO, which relies on rare SNPs to help bolster the DOV of a specific hospital, knowledge of the robustness of the SnAPO method is required. To test this 100 isolates were randomly removed from the Bi-allelic Dataset between 2001 and 2009. This left 832 isolates sampled between 2001 and 2009. This is termed the Robustness Test Subset. SnAPO was recalculated for the 90 isolates sampled in 2010 using this Robustness Test Subset. The random removal of 100 isolates and recalculation of SnAPO was repeated 100 times.

The predicted origin hospital was checked to see if it is the same using both datasets and if the Primary SnAPO result is one of the top three SnAPO results using the Robustness Test Subset. The value of the SnAPO result obtained using the Robustness Test Subset was compared with the value of Primary SnAPO result for each test isolate. Furthermore, the minimum number of hospitals required to obtain 90% of the DOV signal was determined.

6.2.3 Determining the impact of older isolates on SnAPO

It is assumed that the isolates sampled earlier in the dataset would, on average, harbour older SNPs, and these are likely to be in many locations due to the initial rapid dissemination of ST22 across the UK. The SNPs which arose after this spread would more likely be representative of a specific location and so provide more information when attempting to determine the geographic origin of an isolate. This implies that the older isolates may be obscuring some of the signal in SnAPO. Therefore, there may be an age threshold after which excluding the older isolates would provide the cleanest signal and still give the same posited origin hospital in the majority of isolates. The removal of the older isolates, usually containing the more common SNPs, gives further weight to the newer, usually rarer, SNPs. In Chapter 4, SnAPO was used on the Comparison Subset to predict the origin hospital for the 2010 Test Subset isolates. In this chapter all the isolates from the earlier years were iteratively removed,

starting with those sampled in 2001. Each of these smaller subsets of the Comparison Subset will be termed the Attenuated Subsets.

The posited origin hospitals from the Primary SnAPO result was compared to each of the Attenuated Subset results, and it was determined if the Primary SnAPO result is one of the top three Attenuated Subset SnAPO results. The value of the result for each Attenuated Subsets was comapred with the value of the Primary SnAPO result. Furthermore, the minimum number of hospitals required to obtain 90% of the DOV signal was determined. The optimum dataset size will be one which requires the least number of hospitals to provide 90% of the DOV signal while still predicting a hospital which matches the Primary SnAPO result posited origin.

6.2.4 Degrading the SnAPO signal

The effect of modifying the SNPs an isolate expresses on the final posited origin location was explored. To do this the 56 isolates sampled in 2010 which have maximum DOVs higher than 40% were used. For convenience it was assumed that the location posited by SnAPO in Chapter 4 was the correct location. For each isolate simulations were run where an increasing number of SNPs in the target isolate were replaced with randomly selected ones, from the total 5469 possible bi-allelic SNP positions. For each target isolate initially one SNP was replaced, and the new origin location and maximum DOV posited by SnAPO was determined. This simulation was repeated 100 times and observed the number of simulations where the posited origin location is different from that posited by the original SNPs, and the mean maximum DOV. This process was repeated, but instead replaced two SNPs; and so on. This process was repeated until all SNPs that the target isolate expresses were replaced, which for the isolates used ranges between 8 to 83 SNPs. The SNPs that are randomly selected could not be the same as those already expressed by the target isolate.

Therefore, for each target isolate, the number of SNPs that can be replaced before the posited origin location is different from that posited in Chapter 4 was determined. Also, the number of SNPs that can be replaced before the DOV falls below 40% was found. Next, it was determined what the proportion of SNPs that can be replaced in each isolates before either the posited location is different than that previously determined in Chapter 4, or the maximum DOV falls below 40%. This will provide information as to the reliability of each target isolate's posited origin location.

6.3 Results

6.3.1 Comparing the indel data with the Primary SnAPO result

The Indel Dataset predicted the same origin hospital as the Primary SnAPO result in 63 of the 88 test isolates (71.6%), and the Primary SnAPO result appeared in the top three hospitals predicted by the indel data in 85 of the 88 test isolates (96.6%). However, using the Indel Dataset usually (n = 80, 90.9%) gave a lower predicted Diagnostic Origin Value (DOV) for the possible origin hospital than the Primary SnAPO result (Figure 6.1a). A paired Mann-Whitney-U test confirmed that the average DOV for the 88 test isolates as significantly lower ($V = 3830$, $p < 0.001$) when using the indel data than the SNP data (Figure 6.1b). Furthermore, the majority of the test isolates (n = 78, 88.6%) required a greater number of hospitals to explain 90% of the DOV signal observed when using the indel data compared to the SNP data (Figure 6.1c). A paired Mann-Whitney-U test confirmed that the average number of hospitals required to explain 90% of the DOV signal was significantly higher ($V = 206$, $p < 0.001$) when using the indel data than the SNP data (Figure 6.1d).

The Indel Dataset and the Bi-allelic Dataset was combined to determine if this provided a clearer signal for the possible origin of a given isolate. The Combined Dataset predicted the same origin hospital as the Primary SnAPO result in 86 of the 88 test isolates (97.7%), and the Primary SnAPO result appeared in the top three hospitals predicted by the Combined Dataset in all of the test isolates. However, the Combined Dataset usually (n = 83, 94.3%) gave a lower DOV for the possible origin hospital than the SNP-only data (Figure 6.1a). This was confirmed with a paired Mann-Whitney-U test which showed that the average DOV for the Combined Dataset was significantly lower ($V = 3863$, $p < 0.001$) than using just the SNP-only data (Figure 6.1b). Furthermore, all the test isolates (n = 88, 100.0%) required a greater number of hospitals to explain 90% of the DOV signal observed when using the Combined Dataset compared to the SNP-only data (Figure 6.1c). A paired Mann-Whitney-U test confirmed that the average number of hospitals required to explain 90% of the DOV signal was significantly higher ($V = 0$, $p < 0.001$) when using the Combined Dataset than the SNP-only data (Figure 6.1d).

Therefore, it appears that using a SNP-only dataset provides a clearer signal that using either an indel-only dataset or a combined SNP and indel dataset, both in terms of the highest DOV result and the number of hospitals required to obtain 90% of the DOV signal.

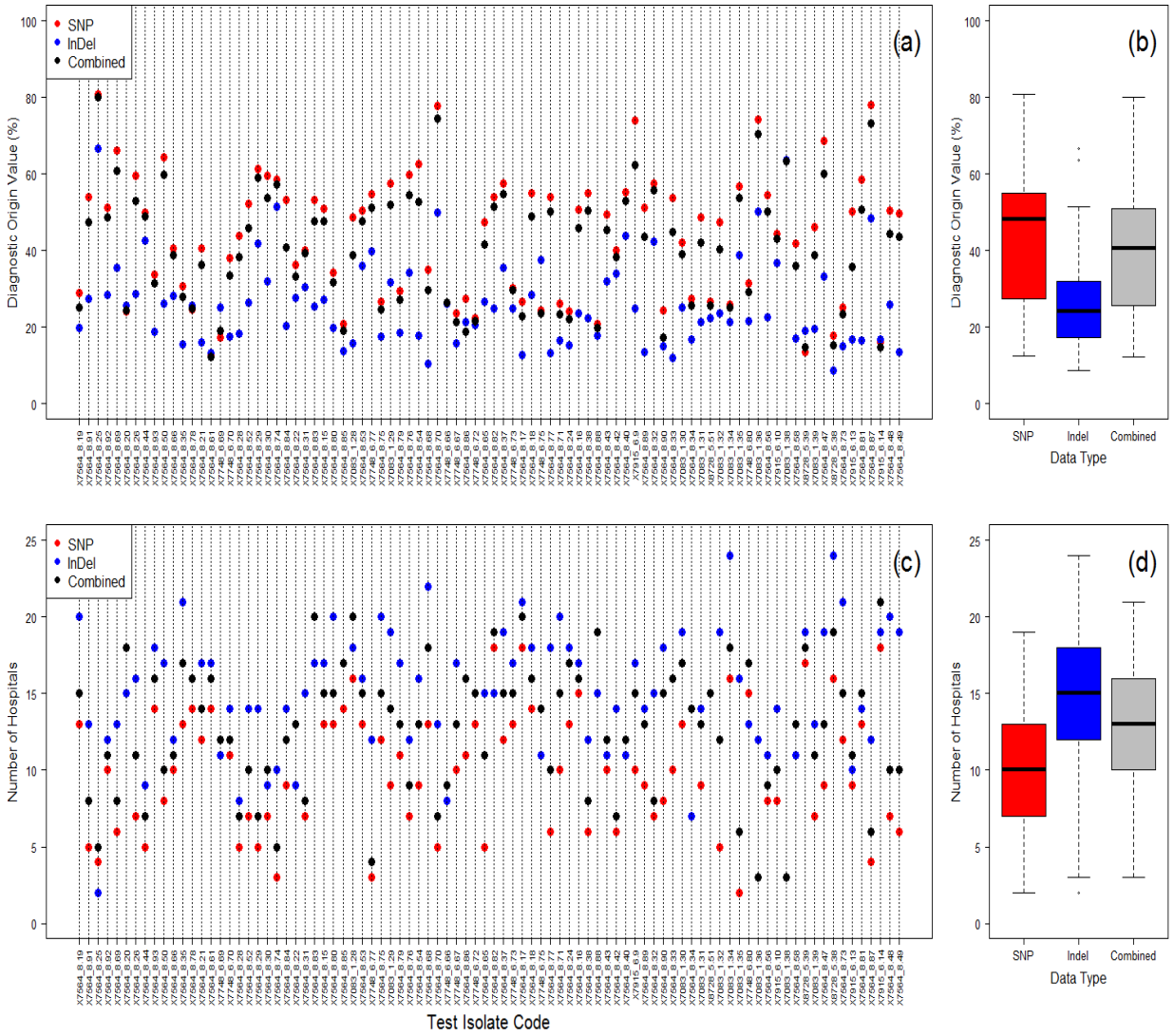


Figure 6.1. The Indel Dataset was processed using a similar method to SnAPO on the 88 isolates in 2010 which harboured indels. For each of the 88 isolates the possible origin hospital was determined using the indel data, the SNP data, and the combined indel and SNP data. For the majority (n = 81) of the 88 test isolates the Diagnostic Origin Value (DOV) for the posited origin hospital is higher using the SNP data than both the indel and the combined data (a). Using a paired Mann-Whitney-U test, the median DOV for the 88 isolates (b) is significantly higher using just the SNP data compared to the indel (V = 3830, p < 0.001) or the combined data (V = 3863, p < 0.001). For the majority (n = 73) of the 88 test isolates the SNP data required fewer hospitals to explain 90% of the DOV signal compared to either the indel or the combined data (c). Using a paired Mann-Whitney-U test, the median number of hospitals required to explain 90% of the DOV signal (d) is significantly smaller when using only the SNP data compared to the indel (V = 206, p < 0.001) or the combined data (V = 0, p < 0.001). The boxplots in (c) and (d) show the median, range, and interquartile range.

6.3.2 The robustness of SnAPO to changes in the dataset

The 832 isolates in the Robustness Test Subset were used to determine the possible origin for each of the 90 2010 Test Subset isolates. The random removal of 100 isolates and the recalculation of SnAPO was repeated 100 times. The average value for the DOV of the posited origin hospital was determined from the 100 replicates for each test isolate (Figure 6.2a), along with the average number of hospitals required to explain 90% of the DOV signal (Figure 6.2b). It was found that the DOV in the Primary SnAPO result fell within the interquartile range of the 100 Robustness Test Subset replicates in 68 (75.5%) of the 2010 Test Subset isolates. Furthermore, the number of hospitals required to predict 90% of the DOV signal in the Primary SnAPO result was within the interquartile range of the 100 replicates in all of the test isolates.

Furthermore, removing 100 isolates resulted in a difference in the posited origin for only a few of the test isolates. Between 83.5% and 98.9% of the test isolates gave the same posited hospital origin as the Primary SnAPO result (Figure 6.2c, red points). There were only 4 replications which had less than 90% of the 90 2010 Test Subset isolates showing the same posited origin as the Primary SnAPO result. However, in each of the 100 replicates the posited origin hospital identified in the Primary SnAPO result was always in the top three posited hospitals using the Robustness Test Subset replicates for each test isolate, bar one (Figure 6.2c, blue points).

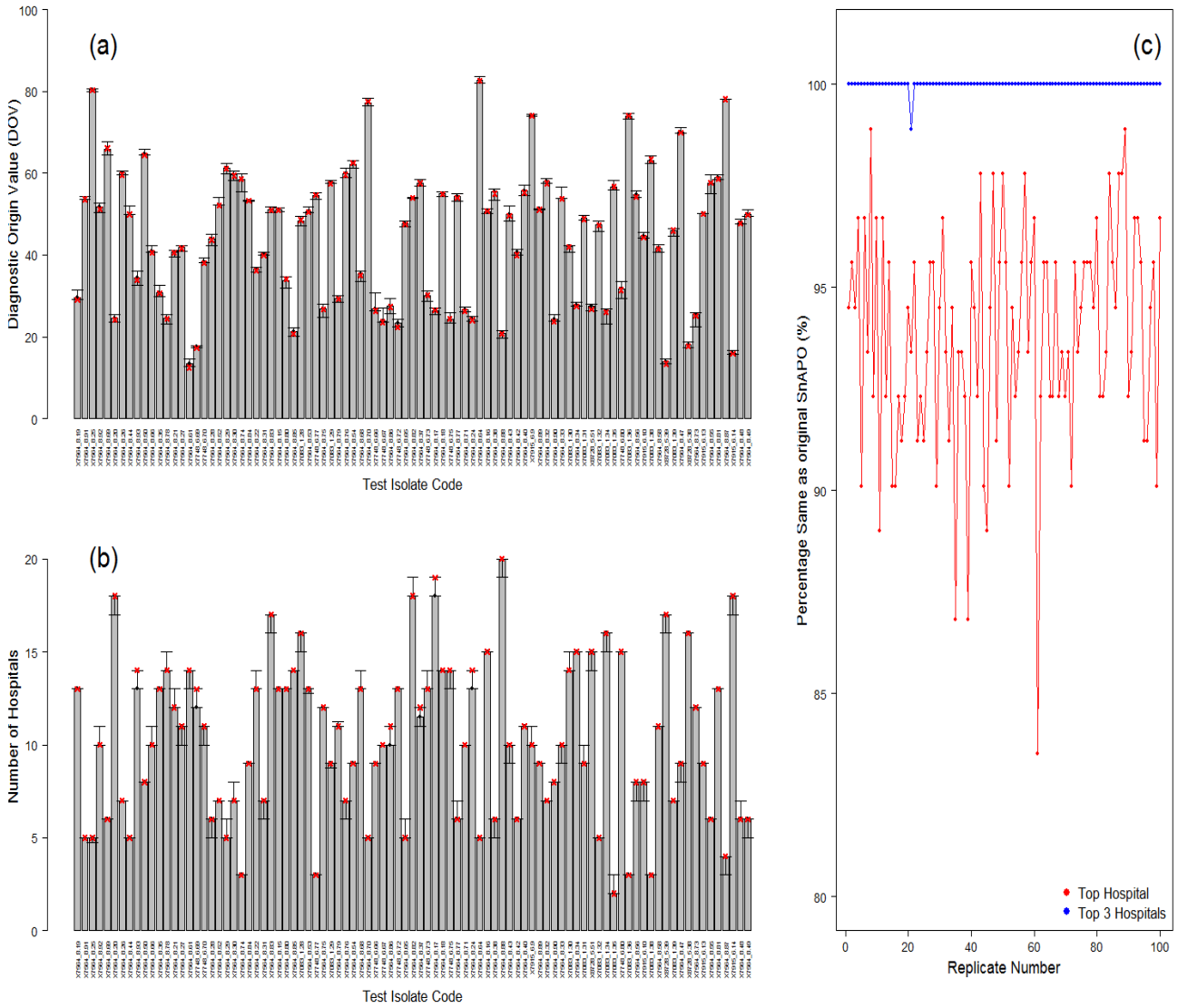


Figure 6.2. The 2010 Test Subset isolates were processed using SnAPO with the Robustness Test Subset isolates. This was repeated 100 times. (a) shows the DOV for the posited origin hospital for each of the 2010 Test Subset isolates. The red points show the value obtained with the Primary SnAPO result, while the grey bars show the median of the 100 replicates of the Robustness Test Subset. The number of hospitals required to explain 90% of the DOV is shown in (b). The red points show the number required using the Bi-allelic Dataset, while the grey bars show the median of the 100 replicates of the Robustness Test Subset. (c) shows the percentage of the 90 test isolates which show the same posited origin hospital as the Primary SnAPO result (red points), and the percentage of isolates where the top posited origin hospital using the Bi-allelic Dataset was one of the top three posited hospitals (blue points) using the Robustness Test Subsets. In (a) and (b) the error bars denote the interquartile range.

6.3.3 The impact of older isolates on SnAPO

As stated in Section 6.2.3, it is possible that older isolates, harbouring older and more common SNPs, may be obscuring some of the possible signal in the SnAPO output. This was tested by iteratively removing older isolates. The Attenuated Subsets are created by iteratively decreasing the size the Comparison Subset. These Attenuated Datasets were used to generate a posited origin hospital for each of the 2010 Test Subset isolates. The Comparison Subset was decreased based on the sampling date, removing one year at a time (Figure 6.3a). Initially the full Comparison Subset (i.e. 2001 to 2009) was used and then each year was removed until a dataset was achieved where only the 2009 isolates were used. It was found that decreasing the Comparison Subset in size changed the posited origin hospital for a number of the test isolates compared to that in the Primary SnAPO result, with a higher discrepancy observed with a smaller dataset (Figure 6.3b, red points). The number of the 90 2010 Test Subset isolates which gave the same posited origin as the Primary SnAPO result ranged from 89 isolates (98.9%) when using isolates sampled from 2002 onwards, down to 41 isolates (45.1%) when only using isolates sampled from 2009 onwards. The posited origin hospital from the Primary SnAPO result remained as one of the top three hospitals with the Attenuated Subsets for all 2010 Test Subset isolates until 2004. However, after 2004 some of the test isolates' top three posited origin hospitals did not contain the Primary SnAPO result. The number of the 90 2010 Test Subset isolates which showed the Primary SnAPO result as one of the top three hospitals (Figure 6.3b, blue points), ranged from 89 isolates (98.9%) when using isolates from 2005 onwards, down to 75 isolates (83.5%) when using only isolates sampled in 2009 onwards. Using only isolates from 2009 onwards showed a noticeable difference from the Primary SnAPO result in the predicted origin hospital.

It was found that there was a slight increase in the average DOV for the test isolates with a decreasing dataset size (Figure 6.3c). However, the individual isolates showed high fluctuations, indicating that these high DOVs could be due to the reduced sampling. Furthermore, the number of hospitals required to explain 90% of the variation showed a decrease with a decreasing dataset size (Figure 6.3d). Once again, the individual isolates showed high fluctuations over the years, indicating the influence of the sampling within a dataset. These two findings appear to indicate that a smaller dataset provides a clearer and sharper signal. However, this must be balanced with the decreasing accuracy of the smaller dataset, as indicated in Figure 6.3b. Therefore, although a clearer signal is obtained with fewer isolates in the dataset, the output is more likely to be erroneous.

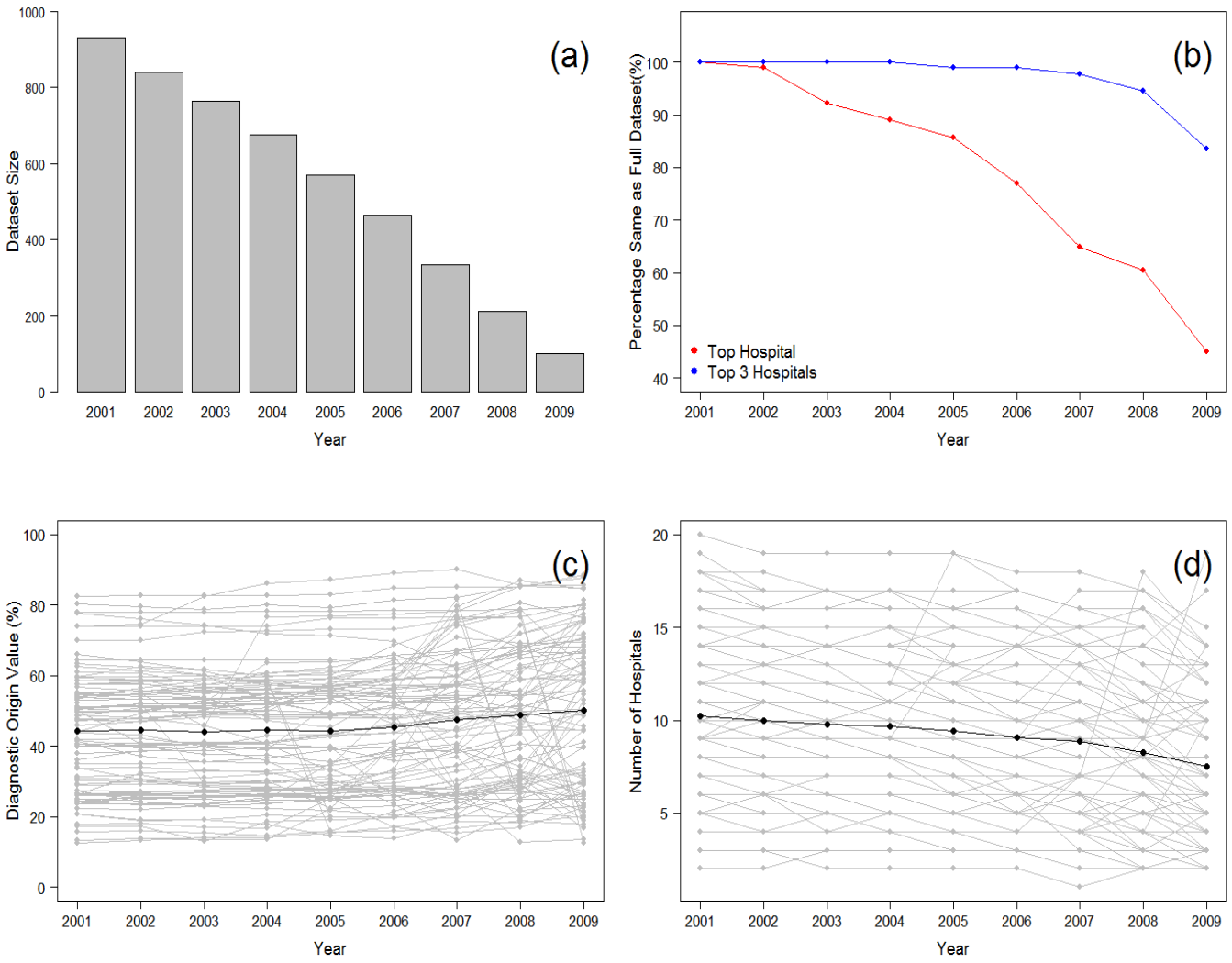


Figure 6.3. The 2010 Test Subset isolates were processed with progressively smaller Attenuated Subsets, iteratively removing the older years. In all figures the x-axis denotes the start year of the Attenuated Subset. (a) shows the decreasing dataset size as each year is removed. (b) shows the percentage of isolates processed using the Attenuated Subsets which show the same posited origin hospital as the Primary SnAPO result (red points), and the percentage which show the origin posited from the Primary SnAPO result as one of the top three hospitals posited. (c) shows the DOV for the posited origin hospital for each isolate (grey points) and the mean value (black points) for the 2010 Test Subset isolates with each Attenuated Subset. (d) shows the number of hospitals required to explain 90% of the DOV for each of the 2010 Test Subset isolates (grey points) and the mean (black points) with each Attenuated Subset.

6.3.4 Degrading the SnAPO signal

The 56 isolates from 2010 which had maximum DOVs higher than 40% were used to test how robust the posited origin location for a target isolate is to changes in the composition of SNPs expressed in the isolate. For each target isolate an increasing number of SNPs were randomly replaced and the new posited location and maximum DOV was noted. This was repeated 100 times for each number of SNPs removed.

As expected, with an increasing number of SNPs replaced there is a decreasing proportion of simulations that posit the same location as was posited in Chapter 4 (Figure 6.4a). Furthermore, there is a decrease in the average maximum DOV when an increasing number of SNPs are replaced (Figure 6.4b).

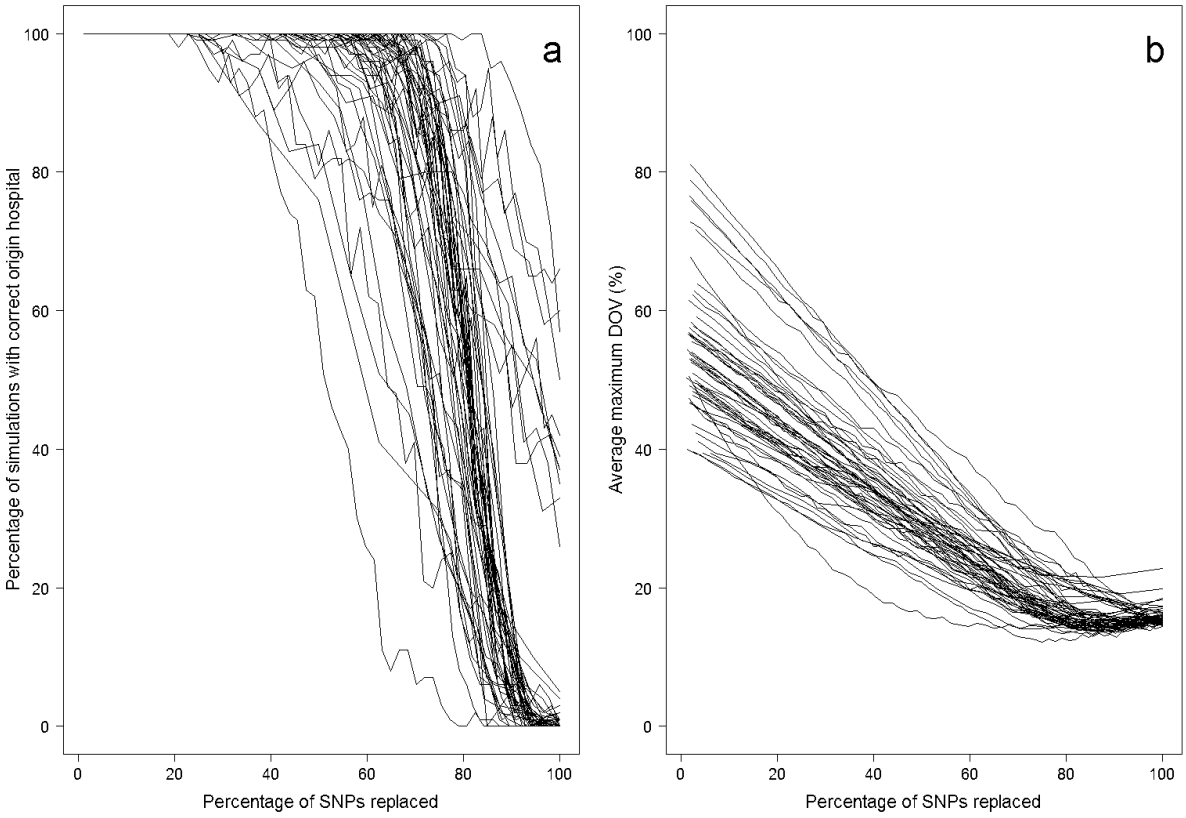


Figure 6.4. For each target isolate, as an increasing number of SNPs are replaced there is a decrease in the number of simulations which show the same hospital as that posited by SnAPO in Chapter 4 (a). Although there is some variation, the majority of isolates appear to show a sharp decline once more than 70% of the SNPs expressed are replaced. There is also a steady decrease in the average maximum DOV for each target isolate with increasing number of SNPs replaced (b). Each black line in a and b corresponds to an individual target isolate.

It was found that, for the majority of isolates, if an average of 29.1% of SNPs were replaced in a target isolate then the output of SnAPO either shows a different posited origin or the maximum DOV falls below 40% (Figure 6.5). If either of these situations arise one cannot be confident of the origin through SnAPO. However, there was considerable range in this output, with some isolates showing a different posited origin hospital or maximum DOV less than 40% once only 2.6% of their SNPs are replaced, while other isolates could withstand 57.1% of the SNPs being replaced. This variation does not appear to correlate with the number of SNPs expressed by an isolate.

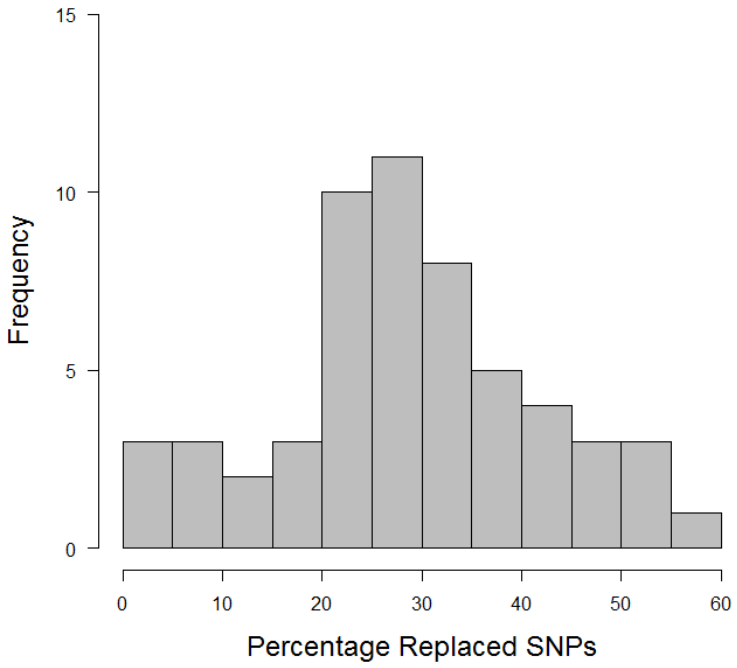


Figure 6.5. If an isolate has more than 29% (mean = 29.1%) of its SNPs replaced then SnAPO either gives a different posited origin from that in Chapter 4 or a maximum DOV less than 40%. However, there is some variation in this value (ranging from 2.6% to 57.1%). This variation does not appear to correlate with the number of SNPs expressed in a target isolate.

6.4 Discussion and Conclusion

I have shown that it is possible to use indels in a similar manner to SnAPO to identify the possible geographic origin of an isolate within the confines of the dataset, either with just the indel data or when combined with the SNP data. Furthermore, the geographic origin posited using the indel data (either on its own or combined with the SNP data) matches the one posited from the SNP-only data in the majority of isolates. However, in the majority of isolates there is a reduction in the clarity of the output DOVs; i.e. the highest DOV is lower and more hospitals are required to explain 90% of the signal seen. These two factors indicate a noisier signal. This is likely due to the reduced number of indels compared to SNPs (899 versus 5469), with a corresponding reduction in number of unique indels per isolate. Furthermore, there is no observable increase in signal clarity for those isolates which had an ambiguous output in the Primary SnAPO result. Therefore, although indels can be used in a similar fashion as SNPs to determine the geographic origin the output is much less clear. It remains to be seen if a similar number of unique indels as SNPs would generate the same output. This would be the obvious next step in comparing indel with SNP data in the SnAPO method, and the successful development of a method using indels to distinguish species (Pereira *et al*., 2010) indicates that this may be a viable avenue of research.

Within the confines of this dataset and the genome of MRSA isolates within a clonal complex, SnAPO appears to be a robust method since a high percentage of the replicates processed with the Robustness Test Dataset showed the same posited origin hospital as the Primary SnAPO result. Furthermore, in all bar one of the replicates the top three possible origin hospitals always contained the hospital which was predicted as the origin in the Primary SnAPO result. Furthermore, removing isolates does not appear to affect the signal clarity and sharpness, indicated by the high number of isolates which showed a DOV within the interquartile range of the replicates, and the number of hospitals required to explain 90% of the signal also falls within the interquartile range of all the replicates. Therefore, not only does SnAPO appear robust in determining where an isolate may have originated from, the reduction in dataset size does not seem to greatly affect the clarity of the signal. It is likely that a great reduction in the dataset, or an increase in the number of hospitals with very few samples, would cause a shift in the signal clarity. However, if sequencing becomes commonplace in healthcare institutions around the UK then this issue would soon be overcome.

It appears that excluding those isolates sampled six years prior to the focal isolate (i.e. pre-2004 for the 2010 Test Subset) might create the optimum dataset. This threshold is based

on increasing the signal clarity (i.e. higher DOVs and a reduced number of hospitals required to predict 90% of the signal) while retaining a high number of the 90 2010 Test Subset isolates predicting the same hospital origin as the Primary SnAPO result (n = 80, 89.0%) and the predicted origin hospital in the Primary SnAPO result is always one of the top three predicted hospitals using the Attenuated Subset sampled from 2004 onwards. Therefore, it does appear that the older isolates are causing increased noise in the final output. If the sequencing of isolates becomes commonplace throughout the UK and SnAPO is utilised I suggest that it might be wise to have a cut-off point in the dataset, before which the isolates are excluded.

Finally, I have also shown, using simulations where the SNPs an isolate expresses were iteratively replaced, that there is some robustness in each isolate for changing the SNPs expressed. To test this the 56 isolates from 2010 which had original SnAPO maximum DOVs higher than 40% were used. As expected, the proportion of simulations which showed the same posited origin as the unaltered SNPs and the average maximum DOV of each simulation showed a decrease with an increasing number of SNPs replaced. However, the maximum DOVs for simulations showed a steady decrease, while there appears to be more of a threshold when it comes to positing the same origin hospital. On average, if an isolate has 29% of its SNPs replaced then one can no longer be confident in the origin location posited, although there is some variation in this distribution. This variation does not appear to correlate with the number of SNPs an isolate expresses, and it is possible this could be due to the uneven sampling in the dataset. It would be interesting, if this dataset is expanded, to re-run these robustness simulations and determine if the variation in robustness has decreased.

In summary, I have explored some of the possible flaws of SnAPO and possible applications to alternative genetic information. I have shown that it is possible to use indel data to generate predicted origin locations, although this is not as clear as using SNP data. I have shown that SnAPO is robust to changes in the database, and that individual isolates vary in their robustness to changes within the SNPs expressed. I have also shown that it is possible that the older SNPs may be obscuring information, and SnAPO might benefit from having a cut-off point.

7 Concluding remarks

The advancements in Whole Genome Sequencing (WGS) have increased the amount of genomic information available for analysis, with current benchtop DNA sequencers able to provide accurate genomic data for tens of isolates within a few days with a low financial cost (Reuter *et al*., 2013). The cost and speed of WGS is decreasing much faster than predicted by Moore's Law (see Section 1.1.1), while the advances in computing power and cost hold consistent with Moore's Law. Furthermore, the speed of WGS has increased so much that culturing the pathogen to acquire sufficient DNA is becoming the limiting step (Köser *et al*., 2014). The advancements in WGS may become an issue, since it cannot be assumed that there will always be sufficient available computing power to implement established analysis techniques on the large databases that would be generated by WGS. For example, it appears that one of the most popular ways to determine the origin of an isolate is to construct a phylogenetic tree (see Section 1.4). However, phylogenetic analyses become computationally prohibitive with larger datasets and are best for retrospective epidemiology. This implies that the implementation of phylogenetic analyses is not practical for rapid results. Therefore, investigation into less computationally demanding methods of analysing the same data to obtain similar conclusions would be an important next step. This thesis describes a proof-of-principle method for evaluating if alternative analysis techniques are viable. In this thesis Methicillin resistant *Staphylococcus aureus* (MRSA) and genomic variation in the form of single nucleotide polymorphisms (SNPs) and indels was used. I move away from the whole genome analysis techniques, such as phylogenetic analysis, and instead focus on individual SNPs. This could be colloquially termed as a "SNP-up" approach. The principles used to develop this "SNP-up" approach would likely be applicable, with some modification, to other pathogens.

MRSA is a major nosocomial infection which has a great impact on healthcare resources (de Angelis *et al*., 2010) and can cause fatalities in patients, especially those which are immunocompromised (Boucher & Corey, 2008; Klevens *et al*., 2007). There is great drive to prevent the dissemination of MRSA, especially since a study by Török *et al*. (2014) concluded that it would be unlikely to achieve zero incidence of MRSA in healthcare institutions. This

implies that the tracking of isolates, determining their likely origin, and prevention of infection could be the more viable avenues to combat MRSA spread by targeting the use of the limited resources and personnel. This is especially important with the increasing evidence that patient referrals are one of the means by which MRSA spreads between healthcare institutions (Ciccolini *et al*., 2013; Donker *et al*., 2010). Additionally, the implementation of WGS in each healthcare institution across the UK, coupled with a centralised online database, would enable the rapid creation of a very large source of MRSA genomes and, as was noted in Reuter *et al* (2013), would also allow real-time monitoring of the geographic spread of a pathogen.

In this thesis I have described the development of a novel heuristic SNP-based Assignment of Pathogen Origin (SnAPO) method which obviates the requirement of a phylogenetic tree by examining the SNPs harboured within an isolate for geographic signal and converting this into a Diagnostic Origin Value (DOV) for each of the locations in the dataset (see Chapter 4). It was the variation in the rarity of the SNPs (see Section 2.4), with some SNPs confined within a geographic location, which indicated that this information could be successfully used to obtain a signal for the possible location of origin of an isolate. SnAPO was developed on a collated ST22 MRSA genome dataset sampled from hospitals across the UK and Ireland between 2001 and 2010. The dataset is one of the largest of its kind, and was collated mainly from the British Society of Antimicrobial Chemotherapy collection, with some genomes taken from the East of England collection. SnAPO has been shown to successfully identify the geographic origin (within the confines of the dataset) of the majority of test isolates. I have shown that SnAPO is more objective than a phylogenetic approach since the output is constant for a given focal isolate and a given dataset, and I also provided a simple and easily interpretable graphical output of the isolate's geographic origin. Furthermore, the high variability of the posited origin locations assigned to the test isolates by each of the three independent investigators exposed the subjectivity of the phylogenetic approach (see Section 4.3.3). Additionally, the use of a Bayesian method with established statistical procedures resulted in a clear signal which concurs with SnAPO in the majority of test isolates. This implies that SnAPO, although a heuristic method, might be a viable one. The methods described in this thesis are applicable to isolates within a clonal complex (CC), and therefore should be part of an analysis pipeline that could include conventional genotyping techniques. In this way it might be possible to focus the limited resources available to combat the dissemination of MRSA.

There is growing evidence that WGS techniques enhance the diagnostic microbiological analyses (Reuter *et al*., 2013). If WGS becomes commonplace throughout

healthcare institutions in the UK, due to the ever decreasing cost, then the implementation of SnAPO and the Bayesian method with a collated database would be a powerful tool for determining the origin of an isolate. The high correlation between the patient referral data and the level of SNP similarity between the MRSA sub-populations (see Section 2.5) indicates that it is the movement of patients which could cause introduction events of MRSA, as posited by Donker *et al*., (2010; 2012; 2014). I showed that it is possible to identify introduction events based on a single signature SNP only ever seen in one location; a Location Specific SNP (LSS). However, this process (see Chapter 3) identified very few introduction events, and still required the use of a phylogenetic tree.

There are clear limitations to SnAPO, even when considering evidence from just this single dataset, with some isolates providing no clear signal of origin and therefore requiring more subjective interpretation by the investigator. One other issue is the likely non-independence of many of the SNPs used to generate the output. In this thesis I have assumed that all SNPs are independent, however as was shown in Section 2.4 this is unlikely to be the case. As discussed in Section 5.4 the non-independence of the SNPs would negatively affect the results of both SnAPO and the Bayesian method by counting the same piece of information multiple times. Further work would be required to clarify this issue. Furthermore, I have assumed that SNPs and indels are inherited stably and are valid indicators of shared heritage in MRSA. This assumption is applicable to the dataset used in this thesis, since it was only comprised of one CC of MRSA. It is possible that this might not be true for all SNPs and indels, with recombination and horizontal gene transfer influencing the posited origin location. Nevertheless, this was a useful dataset with which to develop SnAPO and the Bayesian method. It is important to mention that any hyper-mutator isolates were excluded in this thesis. Due to the assumptions made in SnAPO and the Bayesian approach (see Chapters 4 and 5, respectively), with regards the SNP evolution and inheritance, the origin of hyper-mutators would likely be difficult to resolve. Furthermore, the inclusion of hyper-mutators in the dataset might have a negative impact on the isolates subsequently sampled. Exploration of the impact of hyper-mutators on SnAPO and the Bayesian method would be interesting, though the conservative approach would be to exclude them.

Although the posited origin of an isolate is dependent on the composition of the dataset, I have shown that the SnAPO method is fairly robust to changes in the dataset. I also showed that the older isolates (> 6 years prior to focal isolate) may be obscuring the possible geographic origin of an isolate and so there may be some merit in imposing an age threshold in

the dataset. There is some evidence that the particular clones which comprise the *S. aureus* lineages in circulation go through cycles of expansion and replacement. It is possible that this may be due to the increase in the number of drug resistance measures a clone has that might have a trade-off in cell function and resources; for example, in the case of ST239 which, although a very widespread clone, might be on the decline (Castillo-Ramírez *et al*., 2011). Furthermore, clones within a healthcare institution would compete for resources, and this could be another driving force of the cycles of lineages observed; for example, the replacement of the endemic ST239 by the imported ST22 in hospitals in Singapore from the early 2000s onwards (Hsu *et al*., 2015). This competition may lead to increased genetic diversity within a clone (Hsu *et al*., 2015). Although in this thesis I do suggest imposing an age threshold on the isolates included in the dataset, it is possible that the cycle of MRSA lineages might begin to repeat. If this was to occur it might be informative to once again include the older isolates. Though it is possible that the subsequent cycles will see the SNPs associated with different locations. Additionally, it is possible that some of the isolates in the dataset used in this thesis were obtained from the same patient. Unfortunately, this information is not known for this dataset. I suggest that this might be important information to retain since there is evidence for within-host diversity, though this does appear to be variable over time (Paterson *et al*., 2015). The within-host diversity will likely be less than that within a single hospital. Multiple isolates from a single patient may give a clearer picture of the possible geographic origin of that infection. Additionally, it has been found that non-susceptible MRSA strains can arise from susceptible populations *in vivo* when exposed to antibiotics (van Hal *et al*., 2014). Therefore, it is even more important to ensure that multiple isolates are sequenced from each patient.

SnAPO was created for use on SNP data, but I also have shown that a different source of genetic variation information (i.e. indels) can be used to determine the geographic origin of an isolate. The successful adaptation of SnAPO to other sources of genetic variation indicates that it might be possible to apply this novel method to other systems. However, the use of indels may not be appropriate since it appears to decrease the clarity of the possible geographic origin. It must be noted that it appears that the number of deletions is greater than the number of insertions in eukaryotic organisms (reviewed in Gregory, 2004) and in bacteria it appears that those with the smallest genomes are derived from those with the larger genomes (Mira *et al*., 2001).

In this thesis I have focused on genetic characteristics that are linked to geographic locations. I focused on the hospital and the Referral Cluster (RC) level of geographic resolution. However, a study by Tong *et al*. (2015) showed that there were intra- and inter-ward transmission events within a single hospital. These transmission events occurred at a much finer scale of geographic resolution than that tackled in this thesis. Therefore, it would be interesting to apply the methods described in this thesis to datasets with finer geographic resolution data and investigate if they are able to obtain the same results. Furthermore, other genetic characteristics of MRSA could provide potential avenues of research, such as genetic variation which is connected to virulence factors. Previous work by Laabei *et al*. (2014) has shown that it is possible to predict an isolate's toxicity based solely on the SNPs and indels harboured in it. It should be possible to amend SnAPO for toxicity or virulence characteristics. Since toxicity correlates with severity *in vivo*, this could be a diagnostic tool for rapidly determining the possible health implications of a particular isolate.

Further work must be carried out to determine if the two SNP-based methods for determining an isolate's origin (SnAPO and Bayesian classification) are applicable to other systems. The methods developed in this thesis should be applicable to other CCs of MRSA, due to the clonal nature within a CC, and I suggest that application to a different CC of *Staphylococcus aureus* could be the first step in the further development of these methods. It would be interesting to use these two methods on other datasets, especially ones which have been used to show transmission events. For example, evidence from Holden *et al*. (2013) and Hsu *et al*. (2015) points to the possible introduction of a single isolate from the UK in 2001 as the cause of the ST22 Singapore epidemic. Furthermore, Hsu *et al*. (2015) demonstrated a probable introduction from Singapore to London of the TW20 lineage in 2002. It would be interesting to investigate if the use of SnAPO would corroborate these posited introduction events, when applied to the datasets used in these studies.

If the application of SnAPO and the Bayesian approach to other *S. aureus* lineages is successful, then application to a different species of *Staphylococcus* would be the next step. For example, it has already been discovered, using WGS, that *S. haemolyticus* shows phylogeographic clustering based on the core genome SNPs (Cavanagh *et al*., 2014). I believe that it would be possible to apply SnAPO and the Bayesian method to other species within the *Staphylococcus* genus with a minimal amount of modification. This would be slightly complicated by the variable levels of recombination and HGT observed in different species (e.g. MGEs, see Section 1.2.1).

It would be interesting to investigate if SnAPO and the Bayesian method can be applied to other pathogens. For example, Reuter *et al*. (2013) demonstrated the success of WGS in identifying outbreak isolates in vancomycin-resistant *Enterococcus faecium* and carbapenem-resistant *Enterobacter cloacae*. This study also demonstrated the high concordance between the results of standard clinical microbiological practices and those obtained through WGS. However, in much the same way that high levels of recombination creates issues for the accurate reconstruction of a phylogenetic tree (see Section 1.4.1), the application of SnAPO and the Bayesian method to those pathogens which exhibit high levels of recombination might prove difficult. For example, *Bordetella pertussis* might be a valid target pathogen since it displays a level of recombination only slightly higher than *S. aureus*, while *Streptococcus pneumoniae* would likely not be a valid pathogen for these methods due to the high levels of recombination (Vos & Didelot, 2009). However, this issue need not be insurmountable. It could be possible to identify those regions of the genome which appear to be due to recombination and amend the influence of the SNPs in those regions appropriately. Work by Castillo-Ramírez *et al*. (2012) has shown that it is possible to identify the variable recombination rates between phylogeographic groups of different MRSA ST239 isolates. Furthermore, the construction of SnAPO, which requires multiple SNPs indicating a location to increase the confidence in that location as the posited origin, might mollify the influence of recombination. The application of the two methods developed in this thesis to other non-bacterial pathogens would use the same principles but likely require further development of the method.

In summary, the work in this thesis describes a proof-of-principle method that moves away from the phylogenetic and whole genome analysis techniques and instead uses a small number of genetic signals to obtain a similar result. I showed that genetic signals (such as SNPs and indels) can be utilised in novel ways to rapidly produce a summary of the possible geographic origin of an isolate and has laid the foundation for future work in this direction. The further development of SnAPO and the Bayesian method described in this thesis will be greatly facilitated by the continued decrease in the cost of WGS. The methods developed in this thesis could be added to the suite of analytical epidemiological tools and are a promising indication of the viability of developing cheap, rapid diagnostic tools to be implemented in healthcare institutions. I believe that implementation of WGS and the continued development of SnAPO on the scale proposed could increase the efficiency of the available healthcare resources and personnel in tracking and combating the spread of MRSA. Furthermore, the principles behind

this “SNP-up” approach could have much wider applications than just MRSA, and so further work based on these principles on alternative pathogens could prove to be promising avenues of investigation.

8

Bibliography

Adhikari, R. P., Arvidson, S., & Novick, R. P. (2007). A nonsense mutation in agrA accounts for the defect in agr expression and the avirulence of Staphylococcus aureus 8325-4 traP::kan. *Infection and Immunity*, *75*(9), 4534–40. doi:10.1128/IAI.00679-07

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, *16*(1), 3–14. doi:10.1016/0304-4076(81)90071-3

Akbari, F., & Kjellerup, B. V. (2015). Elimination of Bloodstream Infections Associated with Candida albicans Biofilm in Intravascular Catheters. *Pathogens*, *4*, 457–469. doi:10.3390/pathogens4030457

Albers, C. A., Lunter, G., MacArthur, D. G., McVean, G., Ouwehand, W. H., & Durbin, R. (2011). Dindel: Accurate indel calls from short-read data. *Genome Research*, *21*(6), 961–973. doi:10.1101/gr.112326.110

Ali, S. a, & Hill, D. R. (2003). Giardia intestinalis. *Current Opinion in Infectious Diseases*, *16*(5), 453–60. doi:10.1097/01.qco.0000092817.64370.ab

Allegranzi, B., & Pittet, D. (2009). Role of hand hygiene in healthcare-associated infection prevention. *Journal of Hospital Infection*, *73*(4), 305–315. doi:10.1016/j.jhin.2009.04.019

Amagai, M., Yamaguchi, T., Hanakawa, Y., Nishifuji, K., Sugai, M., & Stanley, J. R. (2002). Staphylococcal Exfoliative Toxin B Specifically Cleaves Desmoglein 1. *The Journal of Investigative Dermatology*, *118*(5), 845–850.

Anderson, S. (1981). Shotgun DNA sequencing using cloned DNase I-generated fragments Nucleic Acids Research. *Nucleic Acids Research*, *9*(13), 3015–3028.

Aubry-Damon, H., Soussy, C. J., & Courvalin, P. (1998). Characterization of mutations in the rpoB gene that confer rifampin resistance in Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, *42*, 2590–2594.

Awadalla, P. (2003). The evolutionary genomics of pathogen recombination. *Nature Reviews. Genetics*, *4*(January), 50–60. doi:10.1038/nrg964

Ayliffe, G. a. (1997). The progressive intercontinental spread of methicillin-resistant Staphylococcus aureus. *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America*, *24 Suppl 1*, S74–S79.

Baba, T., Takeuchi, F., Kuroda, M., Yuzawa, H., Aoki, K., Oguchi, A., & Nagai, Y. (2002). Genome and virulence determinants of high virulence community-acquired MRSA. *Lancet*, *359*, 1819–1827.

Bacaër, N. (2011). Daniel Bernoulli, d'Alembert and the inoculation of smallpox (1760). In *A Short History of Mathematical Population Dynamics* (pp. 21–30). London: Springer London. doi:10.1007/978-0-85729-115-8

Bankston, J. (2004). *Joseph Lister and the Story of Antiseptics (Uncharted, Unexplored, and Unexplained: Scientific Advancements of the 19th Century)*. Mitchell Lane Publishers.

Barreiro, L. B., Laval, G., Quach, H., Patin, E., & Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. *Nature Genetics*, *40*(3), 340–5. doi:10.1038/ng.78

Basic-Hammer, N., Vogel, V., Basset, P., & Blanc, D. S. (2010). Impact of recombination on genetic variability within Staphylococcus aureus clonal complexes. *Infection, Genetics and Evolution*, *10*(7), 1117–23. doi:10.1016/j.meegid.2010.07.013

Bassetti, M., Nicco, E., & Mikulska, M. (2009). Why is community-associated MRSA spreading across the world and how will it change clinical practice? *International Journal of Antimicrobial Agents*, *34 Suppl 1*, S15–9. doi:10.1016/S0924-8579(09)70544-8

Bayes, & Price. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions*, *53*, 370–418.

Bentivoglio, M., & Pacini, P. (1995). Filippo Pacini: A Determined Observer. *Brain Research Bulletin*, *38*(2), 161–165.

Betley, M. J., & Mekalanos, J. J. (1985). Staphylococcal Enterotoxin A Is Encoded by Phage. *Science*, *456*(July), 233–235.

Bogich, T. L., Funk, S., Malcolm, T. R., Chhun, N., Epstein, J. H., Aleksei, A., … Daszak, P. (2013). Using network theory to identify the causes of disease outbreaks of unknown origin. *Journal of the Royal Society INterface*, *10*, 1–9.

Borchers, A. T., Keen, C. L., Huntley, A. C., & Gershwin, M. E. (2014). Lyme disease: A rigorous review of diagnostic criteria and treatment. *Journal of Autoimmunity*, *57*, 82–115. doi:10.1016/j.jaut.2014.09.004

Boucher, H. W., & Corey, G. R. (2008). Epidemiology of methicillin-resistant Staphylococcus aureus. *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America*, *46 Suppl 5*(Suppl 5), S344–9. doi:10.1086/533590

Breiman, R. F., Butler, J. C., Tenover, F. C., Elliott, J. a, & Facklam, R. R. (1994). Emergence of drug-resistant pneumococcal infections in the United States. *JAMA : The Journal of the American Medical Association*, *271*(23), 1831–1835. doi:10.1001/jama.1994.03510470035031

Brossette, S. E., Sprague, A. P., Jones, W. T., & Moser, S. A. (2000). A data mining system for infection control surveillance. *Methods of Information in Medicine*, *39*, 303–310.

Browne, C., Powley, E., Whitehouse, D., Lucas, S., Cowling, P. I., Rohlfshagen, P., … Colton, S. (2012). A Survey of Monte Carlo Tree Search Methods. *IEEE Transactions on Computational Intelligence and AI in Games*, *4*(1), 1–49.

Castanheira, M., Watters, A. A., Mendes, R. E., Farrell, D. J., & Jones, R. N. (2010). Occurrence and molecular characterization of fusidic acid resistance mechanisms among Staphylococcus spp. from European countries (2008). *Journal of Antimicrobial Chemotherapy*, *65*(7), 1353–1358. doi:10.1093/jac/dkq094

Castillo-Ramírez, S., Corander, J., Marttinen, P., Aldeljawi, M., Hanage, W. P., Westh, H., … Feil, E. J. (2012). Phylogeographic variation in recombination rates within a global clone of methicillin-resistant Staphylococcus aureus. *Genome Biology*, *13*(12), R126. Retrieved from http://genomebiology.com/2012/13/12/R126

Castillo-Ramírez, S., Harris, S. R., Holden, M. T. G., He, M., Parkhill, J., Bentley, S. D., & Feil, E. J. (2011). The impact of recombination on dN/dS within recently emerged bacterial clones. *PLoS Pathogens*, *7*(7). doi:10.1371/journal.ppat.1002129

Cavanagh, J. P., Hjerde, E., Holden, M. T. G., Kahlke, T., Klingenberg, C., Flægstad, T., … Sollid, J. U. E. (2014). Whole-genome sequencing reveals clonal expansion of multiresistant Staphylococcus haemolyticus in European hospitals. *The Journal of Antimicrobial Chemotherapy*, *69*(11), 2920–7. doi:10.1093/jac/dku271

Chambers, H. F. (2001). The Changing Epidemiology of Staphylococcus aureus? *Emerging Infectious Diseases*, *7*(2), 178–182.

Chambers, H. F., & Deleo, F. R. (2009). Waves of resistance: Staphylococcus aureus in the antibiotic era. *Nature Reviews. Microbiology*, *7*(9), 629–41. doi:10.1038/nrmicro2200

Chen, H.-J., Hung, W.-C., Tseng, S.-P., Tsai, J.-C., Hsueh, P.-R., & Teng, L.-J. (2010). Fusidic Acid Resistance Determinants in Staphylococcus aureus Clinical Isolates. *Antimicrobial Agents and Chemotherapy*, *54*(12), 4985–4991. doi:10.1128/AAC.00523-10

Chen, J. Q., Wu, Y., Yang, H., Bergelson, J., Kreitman, M., & Tian, D. (2009). Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Molecular Biology and Evolution*, *26*(7), 1523–1531. doi:10.1093/molbev/msp063

Chen, W. P., Hung, C. L., Tsai, S. J., & Lin, Y. L. (2014). Novel and efficient tag SNPs selection algorithms. *Biomed Mater Eng*, *24*(1), 1383–1389. doi:10.3233/BME-130942

Choo, Q. L., Kuo, G., Weiner, a J., Overby, L. R., Bradley, D. W., & Houghton, M. (1989). Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. *Science*, *244*(4902), 359–362. doi:10.1126/science.2523562

Ciccolini, M., Donker, T., Grundmann, H., Bonten, M. J. M., & Woolhouse, M. E. J. (2014). Efficient surveillance for healthcare-associated infections spreading between hospitals.

Proceedings of the National Academy of Sciences of the United States of America, *111*(6), 2271–6. doi:10.1073/pnas.1308062111

Ciccolini, M., Donker, T., Köck, R., Mielke, M., Hendrix, R., Jurke, A., … Friedrich, A. W. (2013). Infection prevention in a connected world: the case for a regional approach. *International Journal of Medical Microbiology : IJMM*, *303*(6-7), 380–7. doi:10.1016/j.ijmm.2013.02.003

Clarke, J., Wu, H., Jayasinghe, L., Patel, A., Reid, S., & Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, *4*, 265–270. doi:10.1038/NNANO.2009.12

Clauditz, A., Resch, A., Wieland, K.-P., Peschel, A., Götz, F., & Gotz, F. (2006). Staphyloxanthin Plays a Role in the Fitness of Staphylococcus aureus and Its Ability To Cope with Oxidative Stress. *Infection and Immunity*, *74*(8), 4950–3. doi:10.1128/IAI.00204-06

Cohn, D. L., Bustreo, F., & Raviglione, M. C. (1997). Drug-Resistant Tuberculosis: Review of the Worldwide Situation and the WHO/IUATLD Global Surveillance Project. *Clinical Infectious Diseases*, *24*(Suppl 1), S121–130.

Cole, A. M., Dewan, P., & Ganz, T. (1999). Innate Antimicrobial Activity of Nasal Secretions. *Infection and Immunity*, *67*(7), 3267–3275.

Cole, A. M., Tahk, S., Oren, A. M. I., Yoshioka, D., Kim, Y., Park, A., & Ganz, T. (2001). Determinants of Staphylococcus aureus Nasal Carriage. *Clinical and Diagnostic Laboratory Immunology*, *8*(6), 1064–1069. doi:10.1128/CDLI.8.6.1064

Colijn, C., & Gardy, J. (2014). Phylogenetic tree shapes resolve disease transmission patterns. *Evolution, Medicine, and Public Health*, *2014*(1), 96–108. doi:10.1093/emph/eou018

Collins, D. W., & Jukes, T. H. (1994). Rates of Transition and Transversion in Coding Sequences since the Human-Rodent Divergence. *Genomics*, *20*, 386–396.

Corander, J., Cui, Y., Koski, T., & Sirén, J. (2013). Have I seen you before? Principles of Bayesian predictive classification revisited. *Statistics and Computing*, *23*(1), 59–73. doi:10.1007/s11222-011-9291-7

Cui, L., Isii, T., Fukuda, M., Ochiai, T., Neoh, H. -m., Camargo, I. L. B. d. C., … Hiramatsu, K. (2010). An RpoB Mutation Confers Dual Heteroresistance to Daptomycin and Vancomycin in Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, *54*(12), 5222–5233. doi:10.1128/AAC.00437-10

Cui, L., Neoh, H. -m., Shoji, M., & Hiramatsu, K. (2009). Contribution of vraSR and graSR Point Mutations to Vancomycin Resistance in Vancomycin-Intermediate Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, *53*(3), 1231–1234. doi:10.1128/AAC.01173-08

Dancer, S. J. (2008). Considering the introduction of universal MRSA screening. *The Journal of Hospital Infection*, *69*(4), 315–20. doi:10.1016/j.jhin.2008.05.002

Dancer, S. J. (2009). The role of environmental cleaning in the control of hospital-acquired infection. *The Journal of Hospital Infection*, *73*(4), 378–85. doi:10.1016/j.jhin.2009.03.030

David, M. Z., & Daum, R. S. (2014). Applying a new technology to an old question: Whole-genome sequencing and staphylococcus aureus acquisition in an intensive care unit. *Clinical Infectious Diseases*, *58*, 619–621. doi:10.1093/cid/cit812

Day, W. H. E. (1987). Computational complexity of inferring phylogenies from dissimilarity matrices. *Bulletin of Mathematical Biology*, *49*(4), 461–467.

De Angelis, G., Murthy, A., Beyersmann, J., & Harbarth, S. (2010). Estimating the impact of healthcare-associated infections on length of stay and costs. *Clinical Microbiology and Infection*, *16*(12), 1729–1735.

Dean, A. G. (1994). Computerizing public health surveillance systems. In S. M. Teutsch & R. E. Churchill (Eds.), *Principles and Practice of Public Health Surveillance* (pp. 200–217). New York: Oxford University Press.

DeLeo, F. R., Kennedy, A. D., Chen, L., Bubeck Wardenburg, J., Kobayashi, S. D., Mathema, B., … Kreiswirth, B. N. (2011). Molecular differentiation of historic phage-type 80/81 and contemporary epidemic Staphylococcus aureus. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(44), 18091–6. doi:10.1073/pnas.1111084108

Delmotte, F., Leterme, N., & Simon, J.-C. (2001). Microsatellite allele sizing: difference between automated capillary electrophoresis and manual technique . *BioTechniques*, *31*, 810–818.

Deurenberg, R. H., & Stobberingh, E. E. (2008). The evolution of Staphylococcus aureus. *Infection, Genetics and Evolution*, *8*(6), 747–63. doi:10.1016/j.meegid.2008.07.007

Deurenberg, R. H., Vink, C., Kalenic, S., Friedrich, A. W., Bruggeman, C. a, & Stobberingh, E. E. (2007). The molecular evolution of methicillin-resistant Staphylococcus aureus. *Clinical Microbiology and Infection*, *13*(3), 222–35. doi:10.1111/j.1469-0691.2006.01573.x

Deurenberg, R. H., Vink, C., Oudhuis, G. J., Mooij, J. E., Driessen, C., Coppens, G., … Stobberingh, E. E. (2005). Different Clonal Complexes of Methicillin-Resistant Staphylococcus aureus Are Disseminated in the Euregio Meuse-Rhine Region. *Antimicrobial Agents and Chemotherapy*, *49*(10), 4263–4271. doi:10.1128/AAC.49.10.4263

Didelot, X. (2010). Sequence-Based Analysis of Bacterial Population Structures. In *Bacterial Population Genetics in Infectious Disease* (pp. 46–47).

Didelot, X., Gardy, J., & Colijn, C. (2014). Bayesian inference of infectious disease transmission from whole-genome sequence data. *Molecular Biology and Evolution*, *31*(7), 1869–79. doi:10.1093/molbev/msu121

Diekema, D. J., Bootsmiller, B. J., Vaughn, T. E., Woolson, R. F., Yankey, J. W., Ernst, E. J., … Doebbeling, B. N. (2004). Antimicrobial Resistance Trends and Outbreak Frequency in United States Hospitals. *Clinical Infectious Diseases*, *38*, 78–85.

Diep, B. A., Stone, G. G., Basuino, L., Graber, C. J., Miller, A., des Etages, S.-A., … Chambers, H. F. (2008). The arginine catabolic mobile element and staphylococcal chromosomal cassette mec linkage: convergence of virulence and resistance in the USA300 clone of methicillin-resistant Staphylococcus aureus. *The Journal of Infectious Diseases*, *197*(11), 1523–30. doi:10.1086/587907

Donker, T., Wallinga, J., & Grundmann, H. (2010). Patient referral patterns and the spread of hospital-acquired infections through national health care networks. *PLoS Computational Biology*, *6*(3), e1000715. doi:10.1371/journal.pcbi.1000715

Donker, T., Wallinga, J., & Grundmann, H. (2014). Dispersal of antibiotic-resistant high-risk clones by hospital networks: changing the patient direction can make all the difference. *The Journal of Hospital Infection*, *86*(1), 34–41. doi:10.1016/j.jhin.2013.06.021

Donker, T., Wallinga, J., Slack, R., & Grundmann, H. (2012). Hospital Networks and the Dispersal of Hospital- Acquired Pathogens by Patient Transfer. *PLoS ONE*, *7*(4), 1–8. doi:10.1371/journal.pone.0035002

Dufour, P., Jarraud, S., Vandenesch, F., Novick, R. P., Bes, M., Etienne, J., & Lina, G. (2002). High Genetic Variability of the agr Locus in Staphylococcus Species. *Journal of Bacteriology*, *184*(4), 1180–1186. doi:10.1128/JB.184.4.1180

Edwards, A. W. F., & Cavalli-Sforza, L. L. (1964). Reconstruction of Evolutionary Trees. *Phenetic and Phylogenetic Classification*, *6*, 67–76.

Efron, B., & Gong, G. (1983). A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. *The American Statistician*, *37*(1), 36–48.

Emonts, M., Uitterlinden, A. G., Nouwen, J. L., Kardys, I., Maat, M. P. M. De, Melles, D. C., … Belkum, A. Van. (2008). Host polymorphisms in interleukin 4, complement factor H, and C-reactive protein associated with nasal carriage of Staphylococcus aureus and occurrence of boils. *The Journal of Infectious Diseases*, *197*(9), 1244–53. doi:10.1086/533501

Enright, M. (2003). The evolution of a resistant pathogen – the case of MRSA. *Current Opinion in Pharmacology*, *3*(5), 474–479. doi:10.1016/S1471-4892(03)00109-7

Enright, M. C., Day, N. P. J., Davies, C. E., & Peacock, S. J. (2000). Multilocus Sequence Typing for Characterization of Methicillin- Resistant and Methicillin-Susceptible Clones of Staphylococcus aureus. *Journal of Clinical Microbiology*, *38*(3), 1008–1015.

Enright, M. C., Robinson, D. A., Randle, G., Feil, E. J., Grundmann, H., & Spratt, B. G. (2002). The evolutionary history of methicillin-resistant Staphylococcus aureus (MRSA). *PNAS*, *99*(11), 7687–7692.

Farris, J. S. (1970). Methods for computing Wagner trees. *Systematic Zoology*, *19*(1), 83–92.

Feil, E. J., Cooper, J. E., Grundmann, H., Robinson, D. A., Enright, M. C., Berendt, T., … Day, N. P. J. (2003). How clonal is Staphylococcus aureus? *Journal of Bacteriology*, *185*(11), 3307–3316. doi:10.1128/JB.185.11.3307

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, *27*(4), 401–410.

Feng, Y., Chen, C.-J., Su, L.-H., Hu, S., Yu, J., & Chiu, C.-H. (2008). Evolution and pathogenesis of Staphylococcus aureus: lessons learned from genotyping and comparative genomics. *FEMS Microbiology Reviews*, *32*(1), 23–37. doi:10.1111/j.1574-6976.2007.00086.x

Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology*, *20*(4), 406–416.

Francois, P., Harbarth, S., Huyghe, A., Renzi, G., Bento, M., Gervaix, A., … Schrenzel, J. (2008). Methicillin-resistant Staphylococcus aureus, Geneva, Switzerland, 1993-2005. *Emerging Infectious Diseases*, *14*(2), 304–307.

Fraser, D. W., Tsai, T. R., Orenstein, W., Parkin, W. E., Beecham, H. J., Sharrar, R. G., … Brachman, P. S. (1977). Legionnaires' Disease: Description of an Epidemic of Pneumonia. *The New England Journal of Medicine*, *297*(22), 1189–1197.

Frénay, H. M. E., Bunschoten, a. E., Schouls, L. M., Leeuwen, W. J., Vandenbroucke-Grauls, C. M. J. E., Verhoef, J., & Mooi, F. R. (1996). Molecular typing of methicillin-resistantStaphylococcus aureus on the basis of protein A gene polymorphism. *European Journal of Clinical Microbiology & Infectious Diseases*, *15*(1), 60–64. doi:10.1007/BF01586186

Fridkin, S. K., Hageman, J. C., Morrison, M., Sanza, L. T., Como-Sabetti, K., Jernigan, J. A., … Farley, M. M. (2005). Methicillin-Resistant Staphylococcus aureus Disease in Three Communities. *The New England Journal of Medicine*, *352*(14), 1436–1444. doi:10.1128/AAC.03622-14

Frost, L. S., Leplae, R., Summers, A. O., & Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nature Reviews. Microbiology*, *3*(9), 722–32. doi:10.1038/nrmicro1235

Fujimura, S., Tokue, Y., & Watanabe, A. (2003). Isoleucyl-tRNA Synthetase Mutations in Staphylococcus aureus Clinical Isolates and In Vitro Selection of Low-Level Mupirocin-Resistant Strains. *Antimicrobial Agents and Chemotherapy*, *47*(10), 3373–3374. doi:10.1128/AAC.47.10.3373

Ganapathy, G., Ramachandran, V., & Warnow, T. (2003). Better Hill-Climbing Searches for Parsimony. In G. Benson & R. Page (Eds.), *Algorithms in Bioinformatics* (Third Inte.). Springer.

Garðarsdóttir, Ó., & Guttormsson, L. (2009). Public health measures against neonatal tetanus on the island of Vestmannaeyjar (Iceland) during the 19th century. *The History of the Family*, *14*(3), 266–279. doi:10.1016/j.hisfam.2009.08.004

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014a). *Bayesian Data Analysis* (Third Edit.). Boca Raton: CRC Press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014b). Example: Informative prior distribution for cancer rates. In *Bayesian Data Analysis* (Third Edit., pp. 47–51). Boca Raton: CRC Press.

Glusman, G., Caballero, J., Mauldin, D. E., Hood, L., & Roach, J. C. (2011). Kaviar: an accessible system for testing SNV novelty. *Bioinformatics*, *27*(22), 3216–7. doi:10.1093/bioinformatics/btr540

Gordon, R. J., & Lowy, F. D. (2008). Pathogenesis of methicillin-resistant Staphylococcus aureus infection. *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America*, *46 Suppl 5*(Suppl 5), S350–9. doi:10.1086/533591

Gouy, M., & Gautier, C. (1982). Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Research*, *10*(22), 7055–7074.

Gregory, T. R. (2004). Insertion–deletion biases and the evolution of genome size. *Gene*, *324*, 15–34. doi:10.1016/j.gene.2003.09.030

Griggs, D. J. (2003). Selection of moxifloxacin-resistant Staphylococcus aureus compared with five other fluoroquinolones. *Journal of Antimicrobial Chemotherapy*, *51*(6), 1403–1407. doi:10.1093/jac/dkg241

Grundmann, H., Aanensen, D. M., van den Wijngaard, C. C., Spratt, B. G., Harmsen, D., & Friedrich, A. W. (2010). Geographic distribution of Staphylococcus aureus causing invasive infections in Europe: a molecular-epidemiological analysis. *PLoS Medicine*, *7*(1), e1000215. doi:10.1371/journal.pmed.1000215

Grundmann, H., Aires-de-Sousa, M., Boyce, J., & Tiemersma, E. (2006). Emergence and resurgence of meticillin-resistant Staphylococcus aureus as a public-health threat. *Lancet*, *368*(9538), 874–85. doi:10.1016/S0140-6736(06)68853-3

Gu, B., Kelesidis, T., Tsiodras, S., Hindler, J., & Humphries, R. M. (2013). The emerging problem of linezolid-resistant Staphylococcus. *Journal of Antimicrobial Chemotherapy*, *68*(1), 4–11. doi:10.1093/jac/dks354

Guglielmini, J., de la Cruz, F., & Rocha, E. P. C. (2013). Evolution of conjugation and type IV secretion systems. *Molecular Biology and Evolution*, *30*(2), 315–31. doi:10.1093/molbev/mss221

Guindon, S., & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, *52*(5), 696–704. doi:10.1080/10635150390235520

Hails, J., Kwaku, F., Wilson, a P., Bellingan, G., & Singer, M. (2003). Large variation in MRSA policies, procedures and prevalence in English intensive care units: a questionnaire analysis. *Intensive Care Medicine*, *29*(3), 481–3. doi:10.1007/s00134-003-1645-y

Hall, N. (2007). Advanced sequencing technologies and their wider impact in microbiology. *The Journal of Experimental Biology*, *209*, 1518–25. doi:10.1242/jeb.001370

Harris, S. R., Cartwright, E. J. P., Török, M. E., Holden, M. T. G., Brown, N. M., Ogilvy-Stuart, A. L., … Peacock, S. J. (2013). Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: A descriptive study. *The Lancet Infectious Diseases*, *13*(2), 130–136. doi:10.1016/S1473-3099(12)70268-2

Harris, S. R., Feil, E. J., Holden, M. T. G., Quail, M. a, Nickerson, E. K., Chantratita, N., … Bentley, S. D. (2010). Evolution of MRSA during hospital transmission and intercontinental spread. *Science (New York, N.Y.)*, *327*(5964), 469–74. doi:10.1126/science.1182395

Harrison, E. M., Paterson, G. K., Holden, M. T. G., Larsen, J., Stegger, M., Larsen, A. R., … Holmes, M. a. (2013). Whole genome sequencing identifies zoonotic transmission of MRSA isolates with the novel mecA homologue mecC. *EMBO Molecular Medicine*, *5*, 509–515. doi:10.1002/emmm.201202413

Harrison, E. M., Weinert, L. A., Holden, M. T. G., Welch, J. J., Wilson, K., Morgan, F. J. E., … Holmes, M. A. (2014). A Shared Population of Epidemic Methicillin-Resistant Staphylococcus aureus 15 Circulates in Humans and Companion Animals. *mBio*, *5*(3), 1–10. doi:10.1128/mBio.00985-13.Invited

Hartman, B. J., & Tomasz, A. (1984). Low-affinity penicillin-binding protein associated with beta-lactam resistance in Staphylococcus aureus. *Journal of Bacteriology*, *158*(2), 513–516.

Hethcote, H. W. (2000). The Mathematics of Infectious Diseases. *Society for Industrial and Applied Mathematics*, *42*(4), 599–653.

Heyde, C. C., Crepel, P., Fienberg, S. E., Seneta, E., & Gani, J. (Eds.). (2001). *Statisticians of the Centuries*. Springer.

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews. Genetics*, *6*(2), 95–108. doi:10.1038/nrg1521

Hogeweg, P., & Hesper, B. (1984). The Alignment of Sets of Sequences and the Construction of Phyletic Trees : An Integrated Method. *Journal of Molecular Evolution*, *20*, 175–186.

Holden, M. T. G., Hsu, L., Kurt, K., Weinert, L. A., Mather, A. E., Harris, S. R., … Nubel, U. (2013). A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant Staphylococcus aureus pandemic. *Genome Research*, *23*(4), 653–64. doi:10.1101/gr.147710.112

Holden, M. T. G., Lindsay, J. a, Corton, C., Quail, M. a, Cockfield, J. D., Pathak, S., … Edgeworth, J. D. (2010). Genome sequence of a recently emerged, highly transmissible, multi-antibiotic- and antiseptic-resistant variant of methicillin-resistant Staphylococcus aureus, sequence type 239 (TW). *Journal of Bacteriology*, *192*(3), 888–92. doi:10.1128/JB.01255-09

Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nature Reviews. Genetics*, *10*(9), 639–50. doi:10.1038/nrg2611

Hon, P. Y., Chan, K. S., Holden, M. T., Harris, S. R., Tan, T. Y., Zu, Y.-B., … Hsu, L. Y. (2013). Arginine catabolic mobile element in methicillin-resistant Staphylococcus aureus (MRSA) clonal group ST239-MRSA-III isolates in Singapore: implications for PCR-based screening tests. *Antimicrobial Agents and Chemotherapy*, *57*(3), 1563–4. doi:10.1128/AAC.02518-12

Howe, R. A., Wootton, M., Noel, A. R., Bowker, K. E., & Walsh, T. R. (2003). Activity of AZD2563, a novel oxazolidinone, against Staphylococcus aureus strains with reduced susceptibility to vancomycin or linezolid. *Antimicrobial Agents and Chemotherapy*, *47*, 3651–3652.

Hsu, L.-Y., Harris, S. R., Chlebowicz, M. a, Lindsay, J. a, Koh, T.-H., Krishnan, P., … Holden, M. T. (2015). Evolutionary dynamics of methicillin-resistant Staphylococcus aureus within a healthcare system. *Genome Biology*, *16*(1), 81. doi:10.1186/s13059-015-0643-z

Hudson, R. R., Slatkint, M., & Wayne, P. (1992). Estimation of Levels of Gene Flow From DNA Sequence Data. *Genetics Society of America*, *589*, 583–589.

Huelsenbeck, J. P., & Bull, J. J. (1996). A likelihood ratio test to detect conflicting phylogenetic signal. *Syst. Biol.*, *45*(1), 92–98.

Hunt, C., Dionne, M., Delorme, M., Murdock, D., Erdrich, A., Wolsey, D., … Danila, R. (1999). Four pediatric deaths from community-acquired methicillin-resistant Staphylococcus aureus - Minnesota and North Dakota, 1997-1999. *JAMA*, *282*(12), 1123–1125.

Hurdle, J. G., O'Neill, A. J., Ingham, E., Fishwick, C., & Chopra, I. (2004). Analysis of Mupirocin Resistance and Fitness in Staphylococcus aureus by Molecular Genetic and Structural Modeling Techniques. *Antimicrobial Agents and Chemotherapy*, *48*(11), 4366–4376. doi:10.1128/AAC.48.11.4366-4376.2004

Hurst, L. D. (2002). The Ka/Ks ratio: diagnosing the form of sequence evolution. *TRENDS in Genetics*, *18*(9), 486. doi:10.1016/S0168-9525(02)02722-1

Huson, D. H., & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, *23*, 254–267. doi:10.1093/molbev/msj030

Huson, D. H., & Scornavacca, C. (2012). Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Systematic Biology*, *61*(6), 1061–1067. doi:10.1093/sysbio/sys062

Ito, T., Kuwahara-Arai, K., Katayama, Y., Uehara, Y., Han, X., Kondo, Y., & Hiramatsu, K. (2013). Methicillin-Resistant Staphylococcus Aureus (MRSA) Protocols. In *Methods in Molecular Biology* (pp. 131–148).

Iwase, T., Uehara, Y., Shinji, H., Tajima, A., Seo, H., Takada, K., … Mizunoe, Y. (2010). Staphylococcus epidermidis Esp inhibits Staphylococcus aureus biofilm formation and nasal colonization. *Nature*, *465*(7296), 346–9. doi:10.1038/nature09074

Jaynes, E. T., & Bretthorst, G. L. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Jensen, L. B., Garcia-Migura, L., Valenzuela, a J. S., Løhr, M., Hasman, H., & Aarestrup, F. M. (2010). A classification system for plasmids from enterococci and other Gram-positive bacteria. *Journal of Microbiological Methods*, *80*(1), 25–43. doi:10.1016/j.mimet.2009.10.012

Jevons, P. (1961). "Celbenin"-resistant staphylococci. *British Medical Journal*, *1*(5219), 113–114.

Johnson, A. P., Aucken, H. M., Cavendish, S., Ganner, M., Wale, M. C. J., Warner, M., … Cookson, B. D. (2001). Dominance of EMRSA-15 and -16 among MRSA causing nosocomial bacteraemia in the UK: analysis of isolates from the European Antimicrobial Resistance Surveillance System (EARSS). *Journal of Antimicrobial Chemotherapy*, *48*, 143–144.

Johnson, A. P., Pearson, A., & Duckworth, G. (2005). Surveillance and epidemiology of MRSA bacteraemia in the UK. *Journal of Antimicrobial Chemotherapy*, *56*(3), 455–462. doi:10.1093/jac/dki266

Johnson, S. (2004). *The Ghost Map: The Story of London's Most Terrifying Epidemic - and How It Changed Science, Cities, and the Modern World*. Riverhead Books.

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, *11*(1), 94. doi:10.1186/1471-2156-11-94

Joshi, G. S., Spontak, J. S., Klapper, D. G., & Richardson, A. R. (2011). Arginine catabolic mobile element encoded speG abrogates the unique hypersensitivity of Staphylococcus aureus to exogenous polyamines. *Molecular Microbiology*, *82*(1), 9–20. doi:10.1111/j.1365-2958.2011.07809.x

Jukes, T. H., & Cantor, C. R. (1969). *Evolution of Protein Molecules*. New York: Academic Press.

Kaito, C., & Sekimizu, K. (2007). Colony spreading in Staphylococcus aureus. *Journal of Bacteriology*, *189*(6), 2553–7. doi:10.1128/JB.01635-06

Kaliner, M. A. (1991). Human Nasal Respiratory Secretions and Host Defense. *American Review of Respiratory Disease*.

Kaneko, J., Kimura, T., Narita, S., Tomita, T., & Kamio, Y. (1998). Complete nucleotide sequence and molecular characterization of the temperate staphylococcal bacteriophage wPVL carrying Panton – Valentine leukocidin genes. *Gene*, *215*, 57–67.

Kanerva, M., Blom, M., Tuominen, U., Kolho, E., Anttila, V.-J., Vaara, M., … Lyytikäinen, O. (2007). Costs of an outbreak of meticillin-resistant Staphylococcus aureus. *The Journal of Hospital Infection*, *66*(1), 22–8. doi:10.1016/j.jhin.2007.02.014

Ke, W., Huang, S. S., Hudson, L. O., Elkins, K. R., Nguyen, C. C., Spratt, B. G., … Lipsitch, M. (2012). Patient sharing and population genetic structure of methicillin-resistant Staphylococcus aureus. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(17), 6763–8. doi:10.1073/pnas.1113578109

Keele, B. F., Heuverswyn, F. Van, Li, Y., Bailes, E., Takehisa, J., Santiago, M. L., … Hahn, B. H. (2006). Chimpanzee Reservoirs of Pandemic and Nonpandemic HIV-1. *Science*, *313*, 523–526.

Khan, A. S., Ksiazek, T. G., & Peters, C. J. (1996). Hantavirus pulmonary syndrome. *The Lancet*, *347*, 739–741.

Khan, S. A., & Novick, R. P. (1983). Complete Nucleotide Sequence of pT181 , a Tetracycline-Resistance Plasmid from Staphyococcus aureus. *Plasmid*, *10*, 251–259.

Khuroo, M. S. (2011). Discovery of hepatitis E: The epidemic non-A, non-B hepatitis 30 years down the memory lane. *Virus Research*, *161*(1), 3–14. doi:10.1016/j.virusres.2011.02.007

Kim, C., Milheiriço, C., Gardete, S., Holmes, M. a, Holden, M. T. G., de Lencastre, H., & Tomasz, A. (2012). Properties of a novel PBP2A protein homolog from Staphylococcus aureus strain LGA251 and its contribution to the β-lactam-resistant phenotype. *The Journal of Biological Chemistry*, *287*(44), 36854–63. doi:10.1074/jbc.M112.395962

Klein, E., Smith, D. L., & Laxminarayan, R. (2007). Hospitalizations and Deaths Caused by Methicillin-Resistant Staphylococcus aureus, United States, 1999-2005. *Emerging Infectious Diseases*, *13*(12), 1999–2005.

Klein, E., Smith, D. L., & Laxminarayan, R. (2009). Community-associated methicillin-resistant Staphylococcus aureus in outpatients, United States, 1999-2006. *Emerging Infectious Diseases*, *15*(12), 1925–30. doi:10.3201/eid1512.081341

Klein, E. Y., Sun, L., Smith, D. L., & Laxminarayan, R. (2013). The changing epidemiology of methicillin-resistant Staphylococcus aureus in the United States: a national observational study. *American Journal of Epidemiology*, *177*(7), 666–74. doi:10.1093/aje/kws273

Klevens, R. M., Morrison, M. A., Nadle, J., Petit, S., Gershman, K., Ray, S., … Fridkin, S. K. (2007). Invasive Methicillin-Resistant Staphylococcus aureus Infections in the United States. *JAMA*, *298*(15), 1763–1771.

Kluytmans, J. A. N., & Verbrugh, H. (1997). Nasal Carriage of Staphylococcus aureus: Epidemiology, Underlying Mechanisms, and Associated Risks. *Clinical Microbiology Reviews*, *10*(3), 505–520.

Köck, R., Schaumburg, F., Mellmann, A., Köksal, M., Jurke, A., Becker, K., & Friedrich, A. W. (2013). Livestock-associated methicillin-resistant Staphylococcus aureus (MRSA) as causes of human infection and colonization in Germany. *PloS One*, *8*(2), e55040. doi:10.1371/journal.pone.0055040

Koh, K. C., Husni, S., Tan, J. E., Tan, C. W., Kunaseelan, S., Nuriah, S., … Morad, Z. (2009). High prevalence of methicillin-resistant Staphylococcus aureus (MRSA) on doctors' neckties. *Med. J. Malaysia2*, *64*(3), 233–235.

Kolaczkowski, B., & Thornton, J. W. (2004). Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*, *431*(October), 980–984. doi:10.1038/nature02917

Köser, C. U., Fraser, L. J., Ioannou, A., Becq, J., Ellington, M. J., Holden, M. T. G., … Peacock, S. J. (2014). Rapid single-colony whole-genome sequencing of bacterial pathogens. *The Journal of Antimicrobial Chemotherapy*, *69*(5), 1275–81. doi:10.1093/jac/dkt494

Köser, C. U., Holden, M. T. G., Ellington, M. J., Cartwright, E. J. P., Brown, N. M., Ogilvy-Stuart, A. L., … Peacock, S. J. (2012). Rapid Whole-Genome Sequencing for Investigation of a Neonatal MRSA Outbreak. *New England Journal of Medicine*, *366*, 2267–2275. doi:10.1056/NEJMoa1109910

Kuhner, M. K., & Felsenstein, J. (1994). A Simulation Comparison of Phylogeny Algorithms under Equal and Unequal Evolutionary Rates. *Molecular Biology and Evolution*, *11*(3), 459–468.

Kuroda, M., Ohta, T., Uchiyama, I., Baba, T., Yuzawa, H., Kobayashi, I., & Cui, L. (2001). Whole genome sequencing of meticillin-resistant Staphylococcus aureus. *Lancet*, *357*, 1225–1240.

Laabei, M., Recker, M., Rudkin, J. K., Aldeljawi, M., Gulay, Z., Sloan, T. J., … Massey, R. C. (2014). Predicting the virulence of MRSA from its genome sequence. *Genome Research*, *24*, 839–849. doi:10.1101/gr.165415.113.Freely

Ladhani, S. (2003). Understanding the mechanism of action of the exfoliative toxins of Staphylococcus aureus. *FEMS Immunology & Medical Microbiology*, *39*(2), 181–189. doi:10.1016/S0928-8244(03)00225-6

Lai, E. (2001). Application of SNP Technologies in Medicine : Lessons Learned and Future Challenges. *Genome*, *11*, 927–929. doi:10.1101/gr.192301.O

Lannergard, J., von Eiff, C., Sander, G., Cordes, T., Seggewiss, J., Peters, G., … Hughes, D. (2008). Identification of the Genetic Basis for Clinical Menadione-Auxotrophic Small-Colony Variant Isolates of Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, *52*(11), 4017–4022. doi:10.1128/AAC.00668-08

Laurent, F., Lelièvre, H., Cornu, M., Vandenesch, F., Carret, G., Etienne, J., & Flandrois, J. (2001). Fitness and competitive growth advantage of new gentamicin-susceptible MRSA clones spreading in French hospitals. *Journal of Antimicrobial Chemotherapy*, *47*, 277–283.

Lee, S. M., Ender, M., Adhikari, R., Smith, J. M. B., Berger-Bächi, B., & Cook, G. M. (2007). Fitness cost of staphylococcal cassette chromosome mec in methicillin-resistant Staphylococcus aureus by way of continuous culture. *Antimicrobial Agents and Chemotherapy*, *51*(4), 1497–9. doi:10.1128/AAC.01239-06

Lee, T.-H., Guo, H., Wang, X., Kim, C., & Paterson, A. H. (2014). SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics*, *15*(1), 162. doi:10.1186/1471-2164-15-162

Letunic, I., & Bork, P. (2006). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, *23*(1), 127–128.

Li, M., Du, X., Villaruz, A. E., Diep, B. A., Wang, D., Song, Y., … Otto, M. (2012). MRSA epidemic linked to a quickly spreading colonization and virulence determinant. *Nature Medicine*, *18*(5), 816–9. doi:10.1038/nm.2692

Licitra, G. (2013). Etymologia: Staphylococcus. *Emerging Infectious Diseases*, *19*(9), 1553.

Lina, G., Pie, Y., Godail-gamot, F., & Peter, M. (1999). Involvement of Panton-Valentine Leukocidin – Producing Staphylococcus aureus in Primary Skin Infections and Pneumonia. *Clinical Infectious Diseases*, *29*, 1128–1132.

Lindsay, J. a, & Holden, M. T. G. (2004). Staphylococcus aureus: superbug, super genome? *Trends in Microbiology*, *12*(8), 378–85. doi:10.1016/j.tim.2004.06.004

Lindsay, J. A. (2014). Staphylococcus aureus genomics and the impact of horizontal gene transfer. *International Journal of Medical Microbiology : IJMM*, *304*(2), 103–9. doi:10.1016/j.ijmm.2013.11.010

Lindsay, J. A., & Holden, M. T. G. (2006). Understanding the rise of the superbug: investigation of the evolution and genomic variation of Staphylococcus aureus. *Functional & Integrative Genomics*, *6*(3), 186–201. doi:10.1007/s10142-005-0019-7

Lindsay, J. A., Knight, G. M., Budd, E. L., & McCarthy, A. J. (2012). Shuffling of mobile genetic elements (MGEs) in successful healthcare-associated MRSA (HA-MRSA). *Mobile Genetic Elements*, (October), 1–5.

Lipsky, B. a., Pecoraro, R. E., Chen, M. S., & Koepsell, T. D. (1987). Factors Affecting Staphylococcal Colonization Among NIDDM Outpatients. *Diabetes Care*, *10*(4), 483–486. doi:10.2337/diacare.10.4.483

Livermore, D. M., Mushtaq, S., Warner, M., & Woodford, N. (2009). Activity of oxazolidinone TR-700 against linezolid-susceptible and -resistant staphylococci and enterococci. *Journal of Antimicrobial Chemotherapy*, *63*(4), 713–715. doi:10.1093/jac/dkp002

Livermore, D. M., Warner, M., Mushtaq, S., North, S., & Woodford, N. (2007). In Vitro Activity of the Oxazolidinone RWJ-416457 against Linezolid-Resistant and -Susceptible Staphylococci and Enterococci. *Antimicrobial Agents and Chemotherapy*, *51*(3), 1112–1114. doi:10.1128/AAC.01347-06

Locke, J. B., Hilgers, M., & Shaw, K. J. (2009). Novel Ribosomal Mutations in Staphylococcus aureus Strains Identified through Selection with the Oxazolidinones Linezolid and Torezolid (TR-700). *Antimicrobial Agents and Chemotherapy*, *53*(12), 5265–5274. doi:10.1128/AAC.00871-09

Lowy, F. D. (1998). Staphylococcus aureus infections. *The New England Journal of Medicine*.

Lu, Y., Goldstein, D. B., Angrist, M., & Cavalleri, G. (2014). Personalized Medicine and Human Genetic Diversity. *Cold Spring Harb Perspect Med*, *4*, 1–12.

Lutz, J. K., van Balen, J., Crawford, J. Mac, Wilkins, J. R., Lee, J., Nava-Hoet, R. C., & Hoet, A. E. (2014). Methicillin-resistant Staphylococcus aureus in public transportation vehicles

(buses): Another piece to the epidemiologic puzzle. *American Journal of Infection Control*, *42*(12), 1285–90. doi:10.1016/j.ajic.2014.08.016

Luzar, M. A., Coles, G. A., Faller, B., Slingeneyer, A., Dah Dah, G., Briat, C., … Peluso, F. (1990). Staphylococcus aureus nasal carriage and infection in patients on continuous ambulatory peritoneal dialysis. *The New England Journal of Medicine*.

Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., … Spratt, B. G. (1998). Multilocus sequence typing : A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. USA*, *95*(March), 3140–3145.

Mantel, N. (1967). The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research*, *27*(2), 209–220.

Marshall, B. J., & Warren, J. R. (1984). Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. *The Lancet*, *1*(June), 1311–5.

Maslow, J. N., Mulligan, M. E., & Arbeit, R. D. (2015). Molecular Epidemiology: Application of Contemporary Techniques to the Typing of Microorganisms. *Clinical Infectious Diseases*, *17*(2), 153–162.

McAdam, P. R., Templeton, K. E., Edwards, G. F., Holden, M. T. G., Feil, E. J., Aanensen, D. M., … Fitzgerald, J. R. (2012). Molecular tracing of the emergence, adaptation, and transmission of hospital-associated methicillin-resistant Staphylococcus aureus. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(23), 9107–12. doi:10.1073/pnas.1202869109

McCarthy, A. J., & Lindsay, J. A. (2012). The distribution of plasmids that carry virulence and resistance genes in Staphylococcus aureus is lineage associated. *BMC Microbiology*.

McCormick, J. K., Yarwood, J. M., & Schlievert, P. M. (2001). Toxic shock syndrome and bacterial superantigens: an update. *Annual Review of Microbiology*, *55*, 77–104.

McDougal, L. K., Steward, C. D., Killgore, G. E., Chaitram, J. M., McAllister, S. K., & Tenover, F. C. (2003). Pulsed-Field Gel Electrophoresis Typing of Oxacillin-Resistant Staphylococcus aureus Isolates from the United States : Establishing a National Database. *Journal of Clinical Microbiology*, *41*(11), 5113–5120. doi:10.1128/JCM.41.11.5113

Mediavilla, J. R., Chen, L., Mathema, B., & Kreiswirth, B. N. (2012). Global epidemiology of community-associated methicillin resistant Staphylococcus aureus (CA-MRSA). *Current Opinion in Microbiology*, *15*(5), 588–95. doi:10.1016/j.mib.2012.08.003

Meirmans, P. G., & Hedrick, P. W. (2011). Assessing population structure: F(ST) and related measures. *Molecular Ecology Resources*, *11*(1), 5–18. doi:10.1111/j.1755-0998.2010.02927.x

Meka, V. G., Pillai, S. K., Sakoulas, G., Wennersten, C., Venkataraman, L., DeGirolami, P. C., … Gold, H. S. (2004). Linezolid Resistance in Sequential Staphylococcus aureus Isolates

Associated with a T2500A Mutation in the 23S rRNA Gene and Loss of a Single Copy of rRNA. *The Journal of Infectious Diseases*, *190*(2), 311–317. doi:10.1086/421471

Mellmann, A., Friedrich, A. W., Rosenkötter, N., Rothgänger, J., Karch, H., Reintjes, R., & Harmsen, D. (2006). Automated DNA sequence-based early warning system for the detection of methicillin-resistant Staphylococcus aureus outbreaks. *PLoS Medicine*, *3*(3), 348–355. doi:10.1371/journal.pmed.0030033

Mendiola, L., González, P., & Cebollada, À. (2015). The relationship between urban development and the environmental impact mobility: A local case study. *Land Use Policy*, *43*, 119–128. doi:10.1016/j.landusepol.2014.11.003

Merson, M. H. (2006). The HIV–AIDS Pandemic at 25 — The Global Response. *The New England Journal of Medicine*, *354*(23), 2414–2417.

Miller, L. G., Perdreau-Remington, F., Rieg, G., Mehdi, S., Perlroth, J., Bayer, A. S., … Spellberg, B. (2005). Necrotizing Fasciitis caused by community-associated methicillin-resistant Staphylococcus aureus in Los Angeles. *The New England Journal of Medicine*, *352*(14), 1445–1453.

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., & Devine, S. E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Research*, *16*, 1182–1190. doi:10.1101/gr.4565806

Mira, A., Ochman, H., & Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes. *TRENDS in Genetics*, *17*(10), 589–596.

Monecke, S., Coombs, G., Shore, A. C., Coleman, D. C., Akpaka, P., Borg, M., … Ehricht, R. (2011). A field guide to pandemic, epidemic and sporadic clones of methicillin-resistant Staphylococcus aureus. *PloS One*, *6*(4), e17936. doi:10.1371/journal.pone.0017936

Morikawa, K., Takemura, A. J., Inose, Y., Tsai, M., Nguyen Thi, L. T., Ohta, T., & Msadek, T. (2012). Expression of a cryptic secondary sigma factor gene unveils natural competence for DNA transformation in Staphylococcus aureus. *PLoS Pathogens*, *8*(11), e1003003. doi:10.1371/journal.ppat.1003003

Morin, P. a., Luikart, G., & Wayne, R. K. (2004). SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, *19*(4), 208–216. doi:10.1016/j.tree.2004.01.009

Naber, C. K. (2009). Staphylococcus aureus bacteremia: epidemiology, pathophysiology, and management strategies. *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America*, *48*(Suppl 4), S231–S237. doi:10.1086/598189

Nachman, M. W. (2001). Single nucleotide polymorphisms and recombination rate in humans. *TRENDS in Genetics*, *17*(9), 481–485.

Neoh, H. -m., Cui, L., Yuzawa, H., Takeuchi, F., Matsuo, M., & Hiramatsu, K. (2008). Mutated Response Regulator graR Is Responsible for Phenotypic Conversion of Staphylococcus aureus from Heterogeneous Vancomycin-Intermediate Resistance to Vancomycin-

Intermediate Resistance. *Antimicrobial Agents and Chemotherapy*, *52*(1), 45–53. doi:10.1128/AAC.00534-07

Newman, M. E. J. (2003). The Structure and Function of Complex Networks. *SIAM Review*, *45*(2), 167. doi:10.1137/S003614450342480

Neyman, J. (1974). Molecular studies: A source of novel statistical problems. In S. S. Gupta & J. Yackel (Eds.), *Statistical Decision Theory and Related Topics* (pp. 1–27). Academic Press.

Nguyen, M. H., Kauffman, C. A., Goodman, R. P., Squier, C., Arbeit, R. D., Singh, N., … Yu, V. L. (1999). Nasal carriage of and infection with Staphylococcus aureus in HIV-infected patients. *Annals of Internal Medicine*, *130*(3), 221–225.

Noble, W. C. (1974). Carriage of Staphylococcus aureus and beta haemolytic Stretococci in relation to race. *Acta Dermatovener*, *54*, 403–405.

North, S. E., Ellington, M. J., Johnson, A. P., Livermore, D. M., & Woodford, N. (2005). Chloramphenicol-selected mutants of Staphylococcus aureus may show cross-resistance to linezolid. In *45th Intersci. Conf. Antimicrob. Agents Chemother.* (pp. C1–1417).

Noto, M. J., Kreiswirth, B. N., Monk, A. B., & Archer, G. L. (2008). Gene acquisition at the insertion site for SCCmec, the genomic island conferring methicillin resistance in Staphylococcus aureus. *Journal of Bacteriology*, *190*(4), 1276–83. doi:10.1128/JB.01128-07

Novick, R. P. (1987). Plasmid Incompatibility. *Microbiology Reviews*, *51*(4), 381–395.

Nucifora, G., Chu, L., Misra, T. K., & Silver, S. (1989). Cadmium resistance from Staphylococcus aureus plasmid pI258 cadA gene results from a cadmium-efflux ATPase. *Proc. Natl. Acad. Sci. USA*, *86*(May), 3544–3548.

Office for National Statistics. (2011). Deaths involving MRSA: England and Wales, 2006 to 2010. *Statistical Bulletin*, (August), 1–13.

Ogston, A. (1882). Micrococcus Poisoning. *Journal of Anatomy and Physiology*, *17*(Pt 1), 24–58.

Olatunde, O. (2013). Deaths involving MRSA: England and Wales, 2008 to 2012. *Office for National Statistics*, (August), 1–12. Retrieved from http://www.ons.gov.uk/ons/dcp171778_317087.pdf

Ometto, L., Stephan, W., & De Lorenzo, D. (2005). Insertion/deletion and nucleotide polymorphism data reveal constraints in Drosophila melanogaster introns and intergenic regions. *Genetics*, *169*(3), 1521–1527. doi:10.1534/genetics.104.037689

Otto, M. (2010). Basis of virulence in community-associated methicillin-resistant Staphylococcus aureus. *Annual Review of Microbiology*, *64*, 143–62. doi:10.1146/annurev.micro.112408.134309

Otto, M. (2013). Community-associated MRSA: what makes them special? *International Journal of Medical Microbiology : IJMM*, *303*(6-7), 324–30. doi:10.1016/j.ijmm.2013.02.007

Patel, M., Thomas, H. C., Room, J., Wilson, Y., Kearns, a, & Gray, J. (2013). Successful control of nosocomial transmission of the USA300 clone of community-acquired meticillin-resistant Staphylococcus aureus in a UK paediatric burns centre. *The Journal of Hospital Infection*, *84*(4), 319–22. doi:10.1016/j.jhin.2013.04.013

Paterson, G. K., Harrison, E. M., Murray, G. G. R., Welch, J. J., Warland, J. H., Holden, M. T. G., … Holmes, M. a. (2015). Capturing the cloud of diversity reveals complexity and heterogeneity of MRSA carriage, infection and transmission. *Nature Communications*, *6*, 6560. doi:10.1038/ncomms7560

Pattengale, N. D., Alipour, M., Bininda-emonds, O. R. P., Moret, B. M. E., & Stamatakis, A. (2010). How Many Bootstrap Replicates Are Necessary? *Journal of Computational Biology*, *17*(3), 337–354.

Patterson, K. D., & Pyle, G. F. (1991). The geography and mortality of the 1918 influenza pandemic. *Bulletin of the History of Medicine*, *65*(1), 4–21.

Peacock, S. J., Justice, A., Griffiths, D., de Silva, G. D. I., Kantzanou, M. N., Crook, D., … Day, N. P. J. (2003). Determinants of Acquisition and Carriage of Staphylococcus aureus in Infancy. *Journal of Clinical Microbiology*, *41*(12), 5718–5725. doi:10.1128/JCM.41.12.5718

Peacock, S. J., Silva, I. De, & Lowy, F. D. (2001). What determines nasal carriage of Staphylococcus aureus? *TRENDS in Microbiology*, *9*(12), 605–610.

Pearson, A., Chronias, A., & Murray, M. (2009). Voluntary and mandatory surveillance for methicillin-resistant Staphylococcus aureus (MRSA) and methicillin-susceptible S. Aureus (MSSA) bacteraemia in England. *Journal of Antimicrobial Chemotherapy*, *64*(SUPPL.1), 11–17. doi:10.1093/jac/dkp260

Pelz, A., Wieland, K.-P., Putzbach, K., Hentschel, P., Albert, K., & Götz, F. (2005). Structure and biosynthesis of staphyloxanthin from Staphylococcus aureus. *The Journal of Biological Chemistry*, *280*(37), 32493–8. doi:10.1074/jbc.M505070200

Pereira, F., Carneiro, J., Matthiesen, R., van Asch, B., Pinto, N., Gusmão, L., & Amorim, A. (2010). Identification of species by multiplex analysis of variable-length sequences. *Nucleic Acids Research*, *38*(22), e203. doi:10.1093/nar/gkq865

Peterson, L. R., & Brossette, S. E. (2002). Hunting Health Care-Associated Infections from the Clinical Microbiology Laboratory : Passive , Active , and Virtual Surveillance. *Journal of Clinical Microbiology*, *40*(1), 1–4. doi:10.1128/JCM.40.1.1

Phillips, M. J., Delsuc, F., & Penny, D. (2004). Genome-scale phylogeny and the detection of systematic biases. *Molecular Biology and Evolution*, *21*(7), 1455–8. doi:10.1093/molbev/msh137

Phillips, S., & Novick, R. P. (1979). Tn554 - a site-specific repressor-controlled transposon in Staphylococcus aureus. *Nature*.

Plinio, P. (1995). *History of Medicine - Volume I: Primitive and Ancient Medicine*. Horatius Press.

Posada, D., & Crandall, K. A. (2001). Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. *PNAS*, *98*(24), 13757–13762.

Prescott, L. M., Harley, J. P., & Klein, D. A. (2005). *Microbiology* (Sixth Edit.). McGraw-Hill.

Price, J. R., Golubchik, T., Cole, K., Wilson, D. J., Crook, D. W., Thwaites, G. E., … Llewelyn, M. J. (2014). Whole-genome sequencing shows that patient-to-patient transmission rarely accounts for acquisition of staphylococcus aureus in an intensive care unit. *Clinical Infectious Diseases*, *58*, 609–618. doi:10.1093/cid/cit807

Price, L. B., Stegger, M., Hasman, H., Aziz, M., Larsen, J., Andersen, S., & Pearson, T. (2012). Adaptation and emergence of Staphylococcus aureus CC39: Host adaptation and emergence of methicillin resistance in livestock. *mBio*, *3*(1), 1–6. doi:10.1128/mBio.00305-11.Editor

Priest, N. K., Rudkin, J. K., Feil, E. J., van den Elsen, J. M. H., Cheung, A., Peacock, S. J., … Massey, R. C. (2012). From genotype to phenotype: can systems biology be used to predict Staphylococcus aureus virulence? *Nature Reviews. Microbiology*, *10*(11), 791–7. doi:10.1038/nrmicro2880

Prunier, A., Malbruny, B., Laurans, M., Brouard, J., Duhamel, J., & Leclercq, R. (2003). High Rate of Macrolide Resistance in Staphylococcus aureus Strains from Patients with Cystic Fibrosis Reveals High Proportions of Hypermutable Strains. *The Journal of Infectious Diseases*, *187*(11), 1709–1716. doi:10.1086/374937

Pynnonen, M., Stephenson, R. E., Schwartz, K., Hernandez, M., & Boles, B. R. (2011). Hemoglobin promotes Staphylococcus aureus nasal colonization. *PLoS Pathogens*, *7*(7), e1002104. doi:10.1371/journal.ppat.1002104

Reagan, D. R., Doebbeling, B. N., Pfaller, M. A., Sheetz, C. T., Houston, A. K., Hollis, R. J., & Wenzel, R. P. (1991). Elimination of coincident Staphylococcus aureus nasal and hand carriage with intranasal application of mupirocin calcium ointment. *Annals of Internal Medicine*, *114*(2), 101–106.

Reuter, S., Ellington, M. J., Cartwright, E. J. P., Köser, C. U., Török, M. E., Gouliouris, T., … Peacock, S. J. (2013). Rapid bacterial whole-genome sequencing to enhance diagnostic and public health microbiology. *JAMA Internal Medicine*, *173*(15), 1397–404. doi:10.1001/jamainternmed.2013.7734

Reuter, S., Török, M. E., Holden, M. T. G., Reynolds, R., Raven, K. E., Blane, B., … Peacock, S. J. (2015). Building a genomic framework for prospective MRSA surveillance in the United Kingdom and the Republic of Ireland. *Genome Research*.

Reyes, J., Rincón, S., Díaz, L., Panesso, D., Contreras, G. a, Zurita, J., … Arias, C. a. (2009). Dissemination of methicillin-resistant Staphylococcus aureus USA300 sequence type 8 lineage in Latin America. *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America*, *49*(12), 1861–7. doi:10.1086/648426

Reynolds, R., Hope, R., & Williams, L. (2008). Survey, laboratory and statistical methods for the BSAC Resistance Surveillance Programmes. *Journal of Antimicrobial Chemotherapy*, *62*(SUPPL. 2), 15–28. doi:10.1093/jac/dkn349

Richards, M. J., Edwards, J. R., Culver, D. H., & Gaynes, R. P. (1999). Nosocomial infections in medical intensive care units in the United States. *Critical Care Medicine*, *27*(5), 887–892. doi:10.1097/00003246-199905000-00020

Richards, M. J., Edwards, J. R., Culver, D. H., & Gaynes, R. P. (2000). Nosocomial Infections in Combined Medical-Surgical Intensive Care Units in the United States. *Infection Control and Hospital Epidemiology*, *21*(8), 510–515.

Riley, L. W., Remis, R. S., Helgerson, S. D., McGee, H. B., Wells, J. G., Davis, B. R., … Cohen, M. L. (1983). Hemorrhagic colitis associated with a rare Escherichia coli serotype. *The New England Journal of Medicine*, *308*(12), 681–685.

Rivas, A. L., Tennenbaum, S. E., Aparicio, J. P., Hoogesteijn, a. L., Mohammed, H. O., Castillo-Chávez, C., & Schwager, S. J. (2003). Critical response time (time available to implement effective measures for epidemic control): Model building and evaluation. *Canadian Journal of Veterinary Research*, *67*(4), 307–311.

Roberts, G. a., Houston, P. J., White, J. H., Chen, K., Stephanou, A. S., Cooper, L. P., … Lindsay, J. a. (2013). Impact of target site distribution for Type i restriction enzymes on the evolution of methicillin-resistant Staphylococcus aureus (MRSA) populations. *Nucleic Acids Research*, *41*(15), 7472–7484. doi:10.1093/nar/gkt535

Roberts, M. C. (2008). Update on macrolide-lincosamide-streptogramin, ketolide, and oxazolidinone resistance genes. *FEMS Microbiology Letters*, *282*(2), 147–159. doi:10.1111/j.1574-6968.2008.01145.x

Robinson, D. A., & Enright, M. C. (2003). Evolutionary Models of the Emergence of Methicillin-Resistant Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, *47*(12), 3926–3934. doi:10.1128/AAC.47.12.3926

Robinson, D. A., Kearns, A. M., Holmes, A., Morrison, D., Grundmann, H., Edwards, G., … Enright, M. C. (2005). Re-emergence of early pandemic Staphylococcus aureus as a community-acquired meticillin-resistant clone. *Lancet*, *365*, 1256–1258.

Rocha, E. P. C., Smith, J. M., Hurst, L. D., Holden, M. T. G., Cooper, J. E., Smith, N. H., & Feil, E. J. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. *Journal of Theoretical Biology*, *239*(2), 226–235. doi:10.1016/j.jtbi.2005.08.037

Roman, R. S., Smith, J., Walker, M., Byrne, S., Ramotar, K., Dyck, B., … Nicolle, L. E. (1997). Rapid geographic spread of a methicillin-resistant Staphylococcus aureus strain. *Clinical*

Infectious Diseases : An Official Publication of the Infectious Diseases Society of America, *25*(3), 698–705. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9314464

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M., & Nyren, P. (1996). Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. *Analytical Biochemistry*, *242*, 84–89.

Ronaghi, M., Uhlen, M., & Nyren, P. (1998). A Sequencing Method Based on Real-Time Pyrophosphate. *Science*, *281*(5375), 363–365. doi:10.1126/science.281.5375.363

Rountree, P. M., & Freeman, B. M. (1955). Infections caused by a particular phage type of Staphylococcus aureus. *The Medical Journal of Australia*, *42*(5), 157–161.

Rowland, S., & Dyke, K. G. H. (1989). Characterization of the staphylococcal beta-lactamase transposon Tn552. *The EMBO Journal*, *8*(9), 2761–2773.

Ruchman, I., & Dodd, K. (1947). The isolation of a strain of Escherichia coli pathogenic for the rabbit's eye from a patient with diarrhea. *Journal of Bacteriology*, *53*(5), 653–656.

Rudkin, J. K., Edwards, A. M., Bowden, M. G., Brown, E. L., Pozzi, C., Waters, E. M., … Massey, R. C. (2012). Methicillin resistance reduces the virulence of healthcare-associated methicillin-resistant Staphylococcus aureus by interfering with the agr quorum sensing system. *The Journal of Infectious Diseases*, *205*(5), 798–806. doi:10.1093/infdis/jir845

Rupp, M. E., Fitzgerald, T., Puumala, S., Anderson, J. R., Craig, R., Iwen, P. C., … Smith, V. (2008). Prospective, controlled, cross-over trial of alcohol-based hand gel in critical care units . *Infection Control and Hospital Epidemiology : The Official Journal of the Society of Hospital Epidemiologists of America*, *29*(1), 8–15. doi:10.1086/524333

Rustchenko-bulgac, E. P., Sherman, F., & Hicks, J. B. (1990). Chromosomal Rearrangements Associated with Morphological Mutants Provide a Means for Genetic Variation of Candida albicans. *Journal of Bacteriology*, *172*(3), 1276–1283.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., & Stein, L. D. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, *409*(February), 928–933.

Saitou, N., & Nei, M. (1987). The Neighbor-joining Method : A New Method for Reconstructing Phylogenetic Trees ’. *Molecular Biology*, *4*(4), 406–425.

Sanches, I. S., Ramirez, M., Troni, H., Abecassis, M., & Padua, M. (1995). Evidence for the Geographic Spread of a Methicillin-Resistant Staphylococcus aureus Clone between Portugal and Spain. *Journal of Clinical Microbiology*, *33*(5), 1243–1246.

Schlievert, P. M., Case, L. C., Nemeth, K. A., Davis, C. C., Sun, Y., Qin, W., … Jones, B. E. (2007). Alpha and beta Chains of Hemoglobin Inhibit Production of Staphylococcus aureus Exotoxins. *Biochemistry*, *46*, 14349–14358.

Schmitt, M., Schuler-Schmid, U., & Schmidt-Lorenz, W. (1990). Temperature limits of growth, TNase and enterotoxin production of Staphylococcus aureus strains isolated from foods. *Internationla Journal of Food Microbiology*, *11*, 1–20.

Schwarz, G. (1978). Estimating the dimenstion of a model. *The Annals of Statistics*, *6*(2), 461–464.

Shittu, A. O., Udo, E. E., & Lin, J. (2007). Insights on Virulence and Antibiotic Resistance: A Review of the Accessory Genome of Staphylococcus aureus. *Wounds*, *19*(9), 237–244.

Shuter, J., Hatcher, V. . ., & Lowy, F. D. (1996). Staphylococcus aureus Binding to Human Nasal Mucin. *Infection and Immunity*, *64*(1), 310–318.

Siepel, A., & Haussler, D. (2004). Phylogenetic Estimation of Context-Dependent Substitution Rates by Maximum Likelihood. *Molecular Biology and Evolution*, *21*, 468–488. doi:10.1093/molbev/msh039

Simoens, S., Ophals, E., & Schuermans, A. (2009). Search and destroy policy for methicillin-resistant Staphylococcus aureus: cost-benefit analysis. *Journal of Advanced Nursing*, *65*(9), 1853–9. doi:10.1111/j.1365-2648.2009.05050.x

Slatkin, M. (2008). Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nature Reviews. Genetics*, *9*(6), 477–485. doi:10.1038/nrg2361

Snow, R. W., Guerra, C. A., Noor, A. M., Myint, H. Y., & Hay, S. I. (2005). The global distribution of clinical episodes of Plasmodium falciparum malaria. *Nature*, *434*, 214–217.

Spoor, L., McAdam, P., Weinert, L. A., Rambaut, A., Hasman, H., Aarestrup, F. M., … Fitzgerald, R. (2013). Livestock origin for a human pandemic clone of community- associated MRSA. *Mbio*, *13*(4), e00356–13. doi:10.1128/mBio.00356-13.Updated

Staden, R. (1979). A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research*, *6*(7), 2601–2610.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. doi:10.1093/bioinformatics/btu033

Steinig, E. J., Andersson, P., Harris, S. R., Sarovich, D. S., Manoharan, A., Coupland, P., … Tong, S. Y. C. (2015). Single-molecule sequencing reveals the molecular basis of multidrug-resistance in ST772 methicillin-resistant Staphylococcus aureus. *BMC Genomics*, *16*, 388. doi:10.1186/s12864-015-1599-9

Sullivan, J., & Joyce, P. (2005). Model Selection in Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *36*(1), 445–466. doi:10.1146/annurev.ecolsys.36.102003.152633

Swaney, S. M., Shinabarger, D. L., Schaadt, R. D., Bock, J. H., Slightom, J. L., & Zurenko, G. E. (1998). Oxazolidinone resistance is associated with a mutation in the peptidyl transferase region of 23S rRNA. In *38th Intersci. Conf. Antimicrob. Agents.* (pp. C–104).

Swofford, D., Olsen, G., Wadell, P., & Hillis, D. M. (1996). Phylogenetic Inference. In Hillis, Moritz, & Mable (Eds.), *Molecular Systematics* (2nd Editio., pp. 407–511). Sinauer.

Tacconelli, E. (2009). Screening and isolation for infection control. *The Journal of Hospital Infection*, *73*(4), 371–7. doi:10.1016/j.jhin.2009.05.002

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, *28*(10), 2731–2739. doi:10.1093/molbev/msr121

Taneike, I., Otsuka, T., Dohmae, S., Saito, K., Ozaki, K., Takano, M., … Yamamoto, T. (2006). Molecular nature of methicillin-resistant Staphylococcus aureus derived from explosive nosocomial outbreaks of the 1980s in Japan. *FEBS Letters*, *580*(9), 2323–34. doi:10.1016/j.febslet.2006.03.049

Thomas, C. M., & Nielsen, K. M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature Reviews. Microbiology*, *3*(9), 711–21. doi:10.1038/nrmicro1234

Thomas, P. E., Klinger, R., Furlong, L. I., Hofmann-Apitius, M., & Friedrich, C. M. (2011). Challenges in the association of human single nucleotide polymorphism mentions with unique database identifiers. *BMC Bioinformatics*, *12 Suppl 4*(Suppl 4), S4. doi:10.1186/1471-2105-12-S4-S4

Todd, J., Elie, F., & Douglas, R. (2005). Epidemiology, Treatment, and Prevention of Community-Acquired Methicillin-resistant Staphylococcus aureus Infections. *Mayo Clinic Proceedings*, *80*(9), 1201–1208.

Todd, J., Fishaut, M., Kapral, F., & Welch, T. (1978). Toxic-shock syndrome associated with phage-group-I Staphylococci. *Lancet*, *2*(8100), 1116–1118. doi:10.1016/S0140-6736(78)92274-2

Tong, S. Y. C., Holden, M. T. G., Nickerson, E. K., Cooper, B. S., Cori, A., Jombart, T., … Peacock, S. J. (2015). Genome sequencing defines phylogeny and spread of methicillin-resistant Staphylococcus aureus in a high transmission setting. *Genome Research*, *25*, 111–118. doi:10.1101/gr.174730.114.Freely

Török, M. E., Harris, S. R., Cartwright, E. J. P., Raven, K. E., Brown, N. M., Allison, M. E. D., … Peacock, S. J. (2014). Zero tolerance for healthcare-associated MRSA bacteraemia: is it realistic? *The Journal of Antimicrobial Chemotherapy*, *69*(8), 2238–45. doi:10.1093/jac/dku128

Tsiodras, S., Gold, H. S., Sakoulas, G., Eliopoulos, G. M., Wennersten, C., Venkataraman, L., … Ferraro, M. J. (2001). Linezolid resistance in a clinical isolate of Staphylococcus aureus. *The Lancet*, *358*(9277), 207–208. doi:10.1016/S0140-6736(01)05410-1

Tsompanidou, E., Sibbald, M. J. J. B., Chlebowicz, M. a, Dreisbach, A., Back, J. W., van Dijl, J. M., … Denham, E. L. (2011). Requirement of the agr locus for colony spreading of

Staphylococcus aureus. *Journal of Bacteriology*, *193*(5), 1267–72. doi:10.1128/JB.01276-10

Tucker, T., Marra, M., & Friedman, J. M. (2009). Massively parallel sequencing: the next big thing in genetic medicine. *American Journal of Human Genetics*, *85*(2), 142–54. doi:10.1016/j.ajhg.2009.06.022

Ubukata, K., Nonoguchi, R., Matsuhashi, M., & Konno, M. (1989). Expression and Inducibility in Staphylococcus aureus of the mecA Gene , Which Encodes a Methicillin-Resistant S . aureus-Specific Penicillin-Binding Protein. *Journal of Bacteriology*, *171*(5), 2882–2885.

Uhlemann, A.-C., Dordel, J., Knox, J. R., Raven, K. E., Parkhill, J., Holden, M. T. G., … Lowy, F. D. (2014). Molecular tracing of the emergence, diversification, and transmission of S. aureus sequence type 8 in a New York community. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(18), 6738–43. doi:10.1073/pnas.1401006111

Väli, U., Brandström, M., Johansson, M., & Ellegren, H. (2008). Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. *BMC Genetics*, *9*, 8. doi:10.1186/1471-2156-9-8

Van Belkum, A., Verkaik, N. J., de Vogel, C. P., Boelens, H. A., Verveer, J., Nouwen, J. L., … Wertheim, H. F. L. (2009). Reclassification of Staphylococcus aureus Nasal Carriage Types. *The Journal of Infectious Diseases*, *199*(12), 1820–1826. doi:10.1086/599119

Van den Akker, E. L. T., Russcher, H., van Rossum, E. F. C., Brinkmann, A. O., de Jong, F. H., Hokken, A., … Lamberts, S. W. J. (2006). Glucocorticoid receptor polymorphism affects transrepression but not transactivation. *The Journal of Clinical Endocrinology and Metabolism*, *91*(7), 2800–3. doi:10.1210/jc.2005-2119

Van Hal, S. J., Steen, J. a, Espedido, B. a, Grimmond, S. M., Cooper, M. a, Holden, M. T. G., … Jensen, S. O. (2014). In vivo evolution of antimicrobial resistance in a series of Staphylococcus aureus patient isolates: the entire picture or a cautionary tale? *The Journal of Antimicrobial Chemotherapy*, *69*(2), 363–7. doi:10.1093/jac/dkt354

Vandendriessche, S., Vanderhaeghen, W., Larsen, J., de Mendonca, R., Hallin, M., Butaye, P., … Denis, O. (2013). High genetic diversity of methicillin-susceptible Staphylococcus aureus (MSSA) from humans and animals on livestock farms and presence of SCCmec remnant DNA in MSSA CC398. *Journal of Antimicrobial Chemotherapy*, *69*(2), 355–362. doi:10.1093/jac/dkt366

Varela, M. a, & Amos, W. (2010). Heterogeneous distribution of SNPs in the human genome: microsatellites as predictors of nucleotide diversity and divergence. *Genomics*, *95*(3), 151–9. doi:10.1016/j.ygeno.2009.12.003

Vester, B., & Douthwaite, S. (2001). Macrolide resistance conferred by base substitutions in 23S rRNA. *Antimicrobial Agents and Chemotherapy*, *45*, 1–12.

Vickers, A. A., Potter, N. J., Fishwick, C. W. G., Chopra, I., & O'Neill, A. J. (2009). Analysis of mutational resistance to trimethoprim in Staphylococcus aureus by genetic and structural

modelling techniques. *Journal of Antimicrobial Chemotherapy*, *63*(6), 1112–1117. doi:10.1093/jac/dkp090

Vignal, A., Milan, D., SanCristobal, M., & Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genet. Sel. Evol*, *34*, 275–305. doi:10.1051/gse

Von Eiff, C., Beckler, K., Machka, K., Stammer, H., & Peters, G. (2001). Nasal carriage as a source of Staphylococcus aureus bacteremia. *New England Journal of Medicine*, *344*(1), 11–16.

Vos, M., & Didelot, X. (2009). A comparison of homologous recombination rates in bacteria and archaea. *The ISME Journal*, *3*(2), 199–208. doi:10.1038/ismej.2008.93

Weinstein, H. J. (1959). The relation between the nasal Staphylococcal carrier state and the incidence of postoperative complciations. *The New England Journal of Medicine*, *260*(26), 1303–1308.

Wertheim, H. F. L., Melles, D. C., Vos, M. C., Leeuwen, W. Van, Belkum, A. Van, Verbrugh, H. A., & Nouwen, J. L. (2005). The role of nasal carriage in Staphylococcus aureus infections. *The Lancet Infectious Diseases*, *5*(December), 751–762.

Wiens, J. J., & Moen, D. S. (2008). Missing data and the accuracy of Bayesian phylogenetics. *Journal of Systematics and Evolution*, *46*(3), 307–314. doi:10.3724/SP.J.1002.2008.08040

Williams, R. E. (1963). Healthy carriage of Staphylococcus aureus: its prevalence and importance. *Bacteriology Review*, *27*(96), 56–71.

Witte, W. (2009). Community-acquired methicillin-resistant Staphylococcus aureus : what do we need to know ? *Clinical Microbiology and Infection*, *15*(Supplement 7), 17–25.

Witte, W., Strommenger, B., Cuny, C., Heuck, D., & Nuebel, U. (2007). Methicillin-resistant Staphylococcus aureus containing the Panton-Valentine leucocidin gene in Germany in 2005 and 2006. *The Journal of Antimicrobial Chemotherapy*, *60*(6), 1258–63. doi:10.1093/jac/dkm384

Woodward, J. (1974). *To do the sick no harm. A study of the British voluntary hospital system to 1875*. (R. Paul & K. Paul, Eds.). London and Boston.

Worby, C. J., Lipsitch, M., & Hanage, W. P. (2014). Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. *PLoS Computational Biology*, *10*(3), e1003549. doi:10.1371/journal.pcbi.1003549

Wu, S., Piscitelli, C., de Lencastre, H., & Tomasz, A. (1996). Tracking the Evolutionary Origin of the Methicillin Resistance Gene: Cloning and Sequencing of a Homologue of mecA from a Methicillin Susceptible Strain of Staphylococcus sciuri. *Microbial Drug Resistance*, *2*(4), 435–441.

Wu, S. W., de Lencastre, H., & Tomasz, A. (2001). Recruitment of the mecA Gene Homologue of Staphylococcus sciuri into a Resistance Determinant and Expression of the Resistant

Phenotype in Staphylococcus aureus. *Journal of Bacteriology*, *183*(8), 2417–2424. doi:10.1128/JB.183.8.2417

Wyklicky, H., & Skopec, M. (1983). Ignaz Philipp Semmelweis, the Prophet of Bacteriology. *Infection Control*, *4*(5), 367–370.

Yamaguchi, T., Hayashi, T., Nakasone, K., Ohnishi, M., Nakayama, K., & Yamada, S. (2000). Phage conversion of exfoliative toxin A production in Staphylococcus aureus. *Molecular Microbiology*, *38*(4), 694–705.

Yang, S.-J., Nast, C. C., Mishra, N. N., Yeaman, M. R., Fey, P. D., & Bayer, A. S. (2010). Cell Wall Thickening Is Not a Universal Accompaniment of the Daptomycin Nonsusceptibility Phenotype in Staphylococcus aureus: Evidence for Multiple Resistance Mechanisms. *Antimicrobial Agents and Chemotherapy*, *54*(8), 3079–3085. doi:10.1128/AAC.00122-10

Yang, Z., & Rannala, B. (1997). Monte Carlo Method A Markov Chain I-I. *Integration The Vlsi Journal*, *14*(7), 717–724. Retrieved from http://mbe.oxfordjournals.org/cgi/content/abstract/14/7/717

Yu, V. L., Goetz, A., Wagener, M., Smith, P. B., Rihs, J. D., Hanchett, J., & Zuravleff, J. J. (1986). Staphylococcus aureus nasal carriage and infection in patients on hemodialysis. *The New England Journal of Medicine*, *315*(2), 91–96.

Appendix A

Supplementary Table A1. The hospitals sampled in this thesis are numbered according to the geographic location and Referral Cluster (RC) they are found in.

No.	Hospital Short Name	Hospital Full Name	City	RC
1	Truro	Treliske Hospital	Truro	2
2	North Devon	North Devon District Hospital	Barnstaple	2
3	Bristol Southmead	Southmead Hospital	Bristol	2
4	Bristol Royal Infirmary	Bristol Royal Infirmary	Bristol	2
5	Southampton	Southampton General Hospital	Southampton	5
6	London Northwick	Northwick Park Hospital	London	1
7	Chelsea	Chelsea & Westminster Hospital	London	1
8	London St. Mary	St. Mary's & Imperial Hospital	London	1
9	Ashford	William Harvey Hospital	Ashford	1
10	London St. Bart's	St. Bart's & The Royal Hospital	London	4
11	UCL	University College Hospital	London	4
12	Chelmsford	Chelmsford Public Health Laboratory	Chelmsford	4
13	Colchester	Colchester Hospital University	Colchester	4
14	Norfolk	Norfolk & Norwich Hospital	Norwich	8
15	West Suffolk	West Suffolk Hospital	Bury St. Edmunds	8
16	Cambridge	Addenbrooke's Hospital	Cambridge	8
17	Papworth	Papworth Hospital	Cambridge	8
18	Leicester	Leicester Royal Infirmary	Leicester	9
19	Nottingham	University Hospital	Nottingham	9
20	Coventry	Coventry & Warwickshire Hospital	Coventry	6
21	Birmingham	Birmingham City Hospital	Birmingham	6
22	Shrewsbury	Royal Shrewsbury Hospital	Shrewsbury	6
23	Chester	Countess of Chester	Chester	7
24	Manchester	Wythenshal Hospital	Manchester	12
25	Sheffield	Northern General Hospital	Sheffield	11
26	York	York Hospitals NHS Trust	York	3
27	Sunderland	Sunderland Royal Hospital	Sunderland	10
28	Newcastle	Freeman Hospital	Newcastle	10
29	Edinburgh Royal Infirmary	Royal Infirmary of Edinburgh	Edinburgh	15
30	Kirkcaldy	Victoria Hospital	Kirkcaldy	15

31	Dundee	Ninewells Hospital	Dundee	15
32	Inverness	Raigmore Hospital	Inverness	15
33	Glasgow South General	Southern General Hospital	Glasgow	15
34	Glasgow Royal Infirmary	Glasgow Royal Infirmary	Glasgow	15
35	Glasgow Victoria	Victoria Infirmary	Glasgow	15
36	Wishaw	Wishaw General Hospital	Wishaw	15
37	Altnaegelvin	Altnaegelvin Area Hospital	Londonderry	14
38	Antrim	Antrim Area Hospital	Antrim	14
39	Belfast	Belfast City Hospital	Belfast	14
40	Ulster	Ulster Hospital	Ulster	14
41	Dublin	Beaumont Hospital	Dublin	13
42	Galway	University College Hospital	Galway	13
43	Cork	Cork University Hospital	Cork	13
44	Bangor	Ysbyty Gwynedd	Bangor	16
45	Wrexham	Wrexham Maelor Hospital	Wrexham	16
46	Cardiff	University Hospital of Wales	Cardiff	16

Supplementary Table A2. The direct geographic distances, in kilometres, between each pairwise hospital in the thesis.

Hospital	Altnaegelvin	Antrim	Dublin	Belfast	Birmingham	Bristol.RI	Cambridge	Chelmsford	Chelsea	Chester	Colchester	Cork	Coventry	Newcastle	Galway	Glasgow.Victoria	Glasgow.RI	Kirkcaldy	Leicester	N.Devon	Sheffield	Dundee	Norfolk	London.Northwick	Papworth	Inverness	Edinburgh.RI	Glasgow.S.Gen	Shrewsbury	Southampton	Bristol.Southmead	London.St.Barts	London.St.Mary	Sunderland	Truro	Ulster	UCL	Cardiff	Nottingham	Ashford	W.Suffolk	Wishaw	Wrexham	Manchester	York	Bangor
Altnaegelvin	0	126	213	157	657	652	883	933	881	527	973	366	710	634	272	349	354	477	733	562	672	506	983	863	854	441	474	343	560	791	650	888	880	654	576	170	884	597	716	1008	941	383	522	583	700	400
Antrim	126	0	149	32.1	535	540	759	810	759	403	848	405	587	512	357	245	250	370	609	468	546	404	857	741	730	376	364	241	439	676	537	765	758	532	509	44.7	761	489	591	885	816	275	400	458	574	282
Dublin	213	149	0	136	488	457	720	765	704	370	809	302	542	545	315	347	352	456	572	351	530	495	832	686	691	505	443	345	389	598	455	712	704	561	368	141	708	397	562	832	782	366	359	437	575	230
Belfast	157	32.1	136	0	503	509	727	778	727	371	817	412	555	485	376	231	236	353	577	441	515	388	826	709	698	374	345	227	407	645	506	733	726	505	488	14.3	729	458	559	853	784	258	368	427	544	250
Birmingham	657	535	488	503	0	136	233	278	224	134	321	733	54.6	282	797	453	454	426	89.7	282	116	457	351	206	203	608	404	460	99.2	182	133	230	223	274	430	491	226	178	96.8	350	295	428	135	108	190	261
Bristol.RI	652	540	457	509	136	0	314	342	269	198	392	658	167	409	746	519	522	522	208	167	253	557	444	253	289	691	499	526	141	142	4.29	278	269	405	307	499	274	66.2	228	392	377	502	182	217	326	261
Cambridge	883	759	720	727	233	314	0	57.3	84.5	357	89.5	961	178	369	1030	636	636	572	150	481	225	590	130	84.1	29.5	761	553	645	332	223	313	77.8	80.9	349	619	714	78.4	377	170	144	63.7	605	363	302	242	492
Chelmsford	933	810	765	778	278	342	57.3	0	78.8	407	50.5	997	224	426	1072	693	693	630	202	508	281	647	126	90.4	83.4	818	611	701	377	231	341	69.2	76.7	406	640	765	72.5	407	225	86.4	57.6	662	412	355	299	538
Chelsea	881	759	704	727	224	269	84.5	78.8	0	358	129	926	175	422	1007	663	663	613	166	432	258	635	201	18.7	85.3	801	593	671	321	152	268	9.67	3.74	404	562	715	6.69	334	197	128	129	634	359	316	294	482
Chester	527	403	370	371	134	198	357	407	358	0	446	639	184	246	685	328	330	325	207	268	162	361	463	339	327	494	302	335	57	301	193	364	357	250	408	358	360	192	193	483	416	307	22.2	70.2	219	140
Colchester	973	848	809	817	321	392	89.5	50.5	129	446	0	1045	267	442	1118	721	721	651	240	558	311	666	86.5	141	119	838	633	729	420	281	391	119	127	420	691	804	123	457	259	85.4	41.6	689	453	391	318	581
Cork	366	405	302	412	733	658	961	997	926	639	1045	0	787	841	165	641	646	756	823	502	801	794	1084	910	933	781	744	638	641	792	658	935	926	856	419	422	931	592	821	1050	1024	664	623	710	856	504
Coventry	710	587	542	555	54.6	167	178	224	175	184	267	787	0	287	852	492	493	454	41.3	325	111	481	297	156	149	639	432	500	154	165	164	180	173	276	471	542	176	219	64.7	298	240	465	188	143	177	314
Newcastle	634	512	545	485	282	409	369	426	422	246	442	841	287	0	852	311	309	214	269	513	176	225	410	406	348	399	200	320	288	453	405	422	418	23.1	654	471	419	427	233	512	401	275	268	196	128	348
Galway	272	357	315	376	797	746	1030	1072	1007	685	1118	165	852	852	0	602	607	727	884	607	846	761	1145	990	1001	712	720	597	699	887	745	1015	1007	869	552	389	1011	681	876	1134	1092	632	673	752	890	545
Glasgow.Victoria	349	245	347	231	453	519	636	693	663	328	721	641	492	311	602	0	5.32	127	498	526	411	158	707	645	609	183	126	8.79	383	628	515	666	661	334	623	219	663	494	469	777	683	36.7	339	349	410	291
Glasgow.RI	354	250	352	236	454	522	636	693	663	330	721	646	493	309	607	5.32	0	123	498	530	411	154	706	645	609	179	123	11.3	385	630	518	667	661	332	628	224	663	497	469	777	682	34.2	341	350	409	295
Kirkcaldy	477	370	456	353	426	522	572	630	613	325	651	756	454	214	727	127	123	0	449	567	355	41.8	624	596	548	188	22.8	134	382	608	518	615	610	237	684	340	612	513	416	716	610	95.2	342	319	333	342
Leicester	733	609	572	577	89.7	208	150	202	166	207	240	823	41.3	269	884	498	498	449	0	366	95.3	473	262	148	121	637	428	506	185	191	205	169	163	255	512	564	165	260	35.5	282	210	469	214	154	149	343
N.Devon	562	468	351	441	282	167	481	508	432	268	558	502	325	513	607	526	530	567	366	0	388	607	610	418	456	708	546	530	228	291	168	442	434	515	147	435	438	106	379	552	544	519	246	321	459	235
Sheffield	672	546	530	515	116	253	225	281	258	162	311	801	111	176	846	411	411	355	95.3	388	0	378	311	241	198	543	335	420	169	277	249	260	255	165	535	502	257	287	61.1	366	275	380	178	93.4	73.5	302
Dundee	506	404	495	388	457	557	590	647	635	361	666	794	481	225	761	158	154	41.8	473	607	378	0	634	619	567	175	62	164	417	638	553	637	633	247	726	376	634	551	439	734	625	130	379	350	349	384
Norfolk	983	857	832	826	351	444	130	126	201	463	86.5	1084	297	410	1145	707	706	624	262	610	311	634	0	207	154	808	608	716	447	350	442	191	198	388	749	812	194	506	270	168	71.3	673	473	400	297	602
London.Northwick	863	741	686	709	206	253	84.1	90.4	18.7	339	141	910	156	406	990	645	645	596	148	418	241	619	207	0	78.4	784	575	653	302	143	253	25.5	17.7	389	550	696	21.4	319	180	146	136	616	341	298	279	464
Papworth	854	730	691	698	203	289	29.5	83.4	85.3	327	119	933	149	348	1001	609	609	548	121	456	198	567	154	78.4	0	737	529	618	302	207	288	81.4	81.6	328	595	685	80.5	352	142	168	91.7	578	334	273	220	462
Inverness	441	376	505	374	608	691	761	818	801	494	838	781	639	399	712	183	179	188	637	708	543	175	808	784	737	0	209	179	551	788	687	803	798	421	805	366	800	671	604	904	798	191	509	500	520	473
Edinburgh.RI	474	364	443	345	404	499	553	611	593	302	633	744	432	200	720	126	123	22.8	428	546	335	62	608	575	529	209	0	134	359	586	495	595	590	223	664	332	592	490	396	697	592	91.2	320	297	315	322
Glasgow.S.Gen	343	241	345	227	460	526	645	701	671	335	729	638	500	320	597	8.79	11.3	134	506	530	420	164	716	653	618	179	134	0	390	636	522	675	669	342	626	216	671	500	478	785	691	45.1	346	357	418	295
Shrewsbury	560	439	389	407	99.2	141	332	377	321	57	420	641	154	288	699	383	385	382	185	228	169	417	447	302	302	551	359	390	0	249	136	327	320	288	372	395	323	141	181	448	393	363	44.5	93.8	236	162
Southampton	791	676	598	645	182	142	223	231	152	301	281	792	165	453	887	628	630	608	191	291	277	638	350	143	207	788	586	636	249	0	143	162	155	441	413	634	159	205	225	262	279	606	293	289	340	395
Bristol.Southmead	650	537	455	506	133	4.29	313	341	268	193	391	658	164	405	745	515	518	518	205	168	249	553	442	253	288	687	495	522	136	143	0	277	269	401	310	496	273	66.3	224	392	376	498	178	213	322	258
London.St.Barts	888	765	712	733	230	278	77.8	69.2	9.67	364	119	935	180	422	1015	666	667	615	169	442	260	637	191	25.5	81.4	803	595	675	327	162	277	0	8.15	404	572	721	4.11	343	199	121	120	637	365	321	294	489
London.St.Mary	880	758	704	726	223	269	80.9	76.7	3.74	357	127	926	173	418	1007	661	661	610	163	434	255	633	198	17.7	81.6	798	590	669	320	155	269	8.15	0	401	564	714	4.22	335	194	128	126	632	358	314	291	481
Sunderland	654	532	561	505	274	405	349	406	404	250	420	856	276	23.1	869	334	332	237	255	515	165	247	388	389	328	421	223	342	288	441	401	404	401	0	658	490	401	426	219	492	379	298	272	195	110	359
Truro	576	509	368	488	430	307	619	640	562	408	691	419	471	654	552	623	628	684	512	147	535	726	749	550	595	805	664	626	372	413	310	572	564	658	0	486	568	252	526	675	681	624	385	466	606	342
Ulster	170	44.7	141	14.3	491	499	714	765	715	358	804	422	542	471	389	219	224	340	564	435	502	376	812	696	685	366	332	216	395	634	496	721	714	490	486	0	717	450	546	841	771	245	356	414	530	240
UCL	884	761	708	729	226	274	78.4	72.5	6.69	360	123	931	176	419	1011	663	663	612	165	438	257	634	194	21.4	80.5	800	592	671	323	159	273	4.11	4.22	401	568	717	0	339	196	125	123	634	361	317	292	485
Cardiff	597	489	397	458	178	66.2	377	407	334	192	457	592	219	427	681	494	497	513	260	106	287	551	506	319	352	671	490	500	141	205	66.3	343	335	426	252	450	339	0	274	458	441	481	172	232	360	218
Nottingham	716	591	562	559	96.8	228	170	225	197	193	259	821	64.7	233	876	469	469	416	35.5	379	61.1	439	270	180	142	604	396	478	181	225	224	199	194	219	526	546	196	274	0	308	225	439	203	133	115	332
Ashford	1008	885	832	853	350	392	144	86.4	128	483	85.4	1050	298	512	1134	777	777	716	282	552	366	734	168	146	168	904	697	785	448	262	392	121	128	492	675	841	125	458	308	0	123	746	485	436	385	609
W.Suffolk	941	816	782	784	295	377	63.7	57.6	129	416	41.6	1024	240	401	1092	683	682	610	210	544	275	625	71.3	136	91.7	798	592	691	393	279	376	120	126	379	681	771	123	441	225	123	0	650	423	358	278	552
Wishaw	383	275	366	258	428	502	605	662	634	307	689	664	465	275	632	36.7	34.2	95.2	469	519	380	130	673	616	578	191	91.2	45.1	363	606	498	637	632	298	624	245	634	481	439	746	650	0	320	322	376	285
Wrexham	522	400	359	368	135	182	363	412	359	22.2	453	623	188	268	673	339	341	342	214	246	178	379	473	341	334	509	320	346	44.5	293	178	365	358	272	385	356	361	172	203	485	423	320	0	88.3	238	129
Manchester	583	458	437	427	108	217	302	355	316	70.2	391	710	143	196	752	349	350	319	154	321	93.4	350	400	298	273	500	297	357	93.8	289	213	321	314	195	466	414	317	232	133	436	358	322	88.3	0	149	209
York	700	574	575	544	190	326	242	299	294	219	318	856	177	128	890	410	409	333	149	459	73.5	349	297	279	220	520	315	418	236	340	322	294	291	110	606	530	292	360	115	385	278	376	238	149	0	352
Bangor	400	282	230	250	261	261	492	538	482	140	581	504	314	348	545	291	295	342	343	235	302	384	602	464	462	473	322	295	162	395	258	489	481	359	342	240	485	218	332	609	552	285	129	209	352	0

Supplementary Table A3. The driving distances, in kilometres, between each pairwise hospital in the thesis. These distances were calculated by using Google Maps to determine the most direct way to drive between each hospital.

Hospital	Altnaegelvin	Antrim	Dublin	Belfast	Birmingham	Bristol.RI	Cambridge	Chelmsford	Chelsea	Chester	Colchester	Cork	Coventry	Newcastle	Galway	Glasgow.Victoria	Glasgow.RI	Kirkcaldy	Leicester	N.Devon	Sheffield	Dundee	Norfolk	London.Northwick	Papworth	Inverness	Edinburgh.RI	Glasgow.S.Gen	Shrewsbury	Southampton	Bristol.Southmead	London.St.Barts	London.St.Mary	Sunderland	Truro	Ulster	UCL	Cardiff	Nottingham	Ashford	W.Suffolk	Wishaw	Wrexham	Manchester	York	Bangor
Altnaegelvin	0	84.3	239	114	615	749	771	837	830	490	856	469	648	445	275	280	285	378	680	911	602	409	799	818	749	551	366	283	521	837	746	817	797	465	1010	121	798	684	641	920	808	306	490	537	543	392
Antrim	84.3	0	189	30.4	576	710	686	798	756	338	722	453	611	361	331	196	200	293	610	870	531	325	715	746	665	467	282	199	471	808	704	763	758	381	964	36.5	762	621	565	867	726	222	448	469	458	342
Dublin	239	189	0	161	361	495	539	602	591	255	650	272	395	526	218	361	365	458	427	684	371	489	607	577	541	632	447	363	286	602	505	590	561	533	757	169	563	449	405	690	576	386	263	302	444	157
Belfast	114	30.4	161	0	536	652	694	741	760	412	777	424	570	366	323	200	208	300	602	819	525	329	719	748	669	474	289	206	443	740	676	747	718	386	932	10.2	721	606	555	504	715	229	420	459	461	314
Birmingham	615	576	361	536	0	139	167	230	192	121	252	667	34.9	335	611	461	461	528	72.7	284	140	569	255	179	143	731	470	468	77.4	213	133	194	189	324	400	570	191	168	84.8	284	204	441	111	123	214	208
Bristol.RI	749	710	495	652	139	0	268	251	188	239	307	558	165	476	621	596	596	662	195	153	284	704	341	189	229	865	604	602	172	120	6	192	190	464	271	697	192	67.4	225	280	314	576	239	257	362	342
Cambridge	771	686	539	694	167	268	0	81.1	96	281	73.4	806	134	378	770	571	572	594	123	401	205	678	103	85.1	23.2	797	539	578	238	205	264	90.8	93.5	366	557	699	92.6	318	152	165	47.7	551	284	259	263	367
Chelmsford	837	798	602	741	230	251	81.1	0	65.6	343	41.3	889	192	450	828	629	643	666	195	406	277	737	132	80.7	84	869	611	650	300	187	268	68.3	56	439	509	757	53.5	301	224	106	72.8	623	335	345	322	431
Chelsea	830	756	591	760	192	188	96	65.6	0	319	101	850	155	450	804	640	640	708	163	322	268	748	188	19	99.2	891	638	649	262	123	186	6.3	5.5	437	427	768	5.8	238	213	93.3	140	620	310	309	334	420
Chester	490	338	255	412	121	239	281	343	319	0	365	523	155	300	469	386	381	448	167	394	128	495	369	306	257	651	390	388	69.8	350	249	320	309	289	523	428	312	329	154	446	318	361	22.1	58.3	187	101
Colchester	856	722	650	777	252	307	73.4	41.3	101	365	0	903	215	467	857	657	657	725	205	442	290	764	94.8	117	102	907	653	666	323	234	305	105	108	453	547	785	105	357	244	153	47.8	636	370	359	349	482
Cork	469	453	272	424	667	558	806	889	850	523	903	0	680	789	199	623	629	722	712	683	639	753	859	843	794	895	711	627	553	869	541	722	829	801	819	433	742	494	673	958	844	650	516	570	689	425
Coventry	648	611	395	570	34.9	165	134	192	155	155	215	680	0	331	620	497	496	563	40.4	298	145	606	218	142	106	753	495	503	113	198	147	156	154	320	427	606	156	206	89.9	261	172	476	148	156	207	242
Newcastle	445	361	526	366	335	476	378	450	450	300	467	789	331	0	732	244	232	218	303	630	209	270	409	440	358	421	163	243	365	519	474	451	443	24.2	772	372	445	516	258	535	415	207	321	247	139	391
Galway	275	331	218	323	611	621	770	828	804	469	857	199	620	732	0	568	573	667	650	770	585	698	813	797	748	841	653	571	526	834	615	812	806	746	894	337	808	587	638	918	811	588	470	518	643	369
Glasgow.Victoria	280	196	361	200	461	596	571	629	640	386	657	623	497	244	568	0	5.9	94.7	494	741	402	134	599	630	548	277	84.4	7.5	450	685	590	649	637	264	860	208	637	625	442	769	612	27.2	402	353	341	476
Glasgow.RI	285	200	365	208	461	596	572	643	640	381	657	629	496	232	573	5.9	0	81.4	499	741	395	122	599	629	548	267	82.6	10.1	450	684	589	648	643	252	857	213	646	636	452	752	609	25.8	402	353	340	472
Kirkcaldy	378	293	458	300	528	662	594	666	708	448	725	722	563	218	667	94.7	81.4	0	518	807	425	52.6	667	697	616	238	51.9	92	511	752	657	722	669	238	924	307	713	703	473	778	628	83.8	470	421	376	538
Leicester	680	610	427	602	72.7	195	123	195	163	167	205	712	40.4	303	650	494	499	518	0	340	111	606	184	149	96.1	724	466	532	133	234	181	164	160	291	468	617	163	235	50.4	254	163	485	171	160	191	265
N.Devon	911	870	684	819	284	153	401	406	322	394	442	683	298	630	770	741	741	807	340	0	432	849	516	316	382	992	746	749	327	227	159	326	322	612	142	854	326	203	376	404	468	721	391	402	508	448
Sheffield	602	531	371	525	140	284	205	277	268	128	290	639	145	209	585	402	395	425	111	432	0	511	232	254	181	628	370	409	187	345	281	267	265	197	549	531	267	324	64	360	242	383	145	72.1	95	218
Dundee	409	325	489	329	569	704	678	737	748	495	764	753	606	270	698	134	122	52.6	606	849	511	0	707	738	656	205	109	131	558	794	698	763	745	297	964	336	745	733	550	836	720	135	511	461	427	584
Norfolk	799	715	607	719	255	341	103	132	188	369	94.8	859	218	409	813	599	599	667	184	516	232	707	0	183	116	850	596	607	326	318	338	177	182	395	620	727	179	390	193	252	66	579	359	301	273	447
London.Northwick	818	746	577	748	179	189	85.1	80.7	19	306	117	843	142	440	797	630	629	697	149	316	254	738	183	0	89.2	882	628	639	252	120	181	23.5	17.4	427	422	759	20.2	232	200	145	136	610	299	297	323	419
Papworth	749	665	541	669	143	229	23.2	84	99.2	257	102	794	106	358	748	548	548	616	96.1	382	181	656	116	89.2	0	799	546	557	215	207	227	99.5	95.9	343	502	677	96.3	276	129	183	68.1	528	249	251	241	382
Inverness	551	467	632	474	731	865	797	869	891	651	907	895	753	421	841	277	267	238	724	992	628	205	850	882	799	0	264	270	715	956	840	905	862	442	1127	479	876	906	676	981	837	278	654	624	579	742
Edinburgh.RI	366	282	447	289	470	604	539	611	638	390	653	711	495	163	653	84.4	82.6	51.9	466	746	370	109	596	628	546	264	0	93	453	694	598	647	614	183	866	295	618	644	418	694	578	71.7	411	362	324	480
Glasgow.S.Gen	283	199	363	206	468	602	578	650	649	388	666	627	503	243	571	7.5	10.1	92	532	749	409	131	607	639	557	270	93	0	451	692	595	654	650	263	864	210	653	643	457	759	617	32.6	408	360	346	478
Shrewsbury	521	471	286	443	77.4	172	238	300	262	69.8	323	553	113	365	526	450	450	511	133	327	187	558	326	252	215	715	453	451	0	301	178	264	259	353	430	491	261	176	147	369	275	425	49.9	107	238	132
Southampton	837	808	602	740	213	120	205	187	123	350	234	869	198	519	834	685	684	752	234	227	345	794	318	120	207	956	694	692	301	0	134	129	129	507	320	799	130	185	289	199	251	666	336	346	414	431
Bristol.Southmead	746	704	505	676	133	6	264	268	186	249	305	541	147	474	615	590	589	657	181	159	281	698	338	181	227	840	598	595	178	134	0	191	186	458	279	703	192	61.1	225	278	314	569	225	251	356	348
London.St.Barts	817	763	590	747	194	192	90.8	68.3	6.3	320	105	722	156	451	812	649	648	722	164	326	267	763	177	23.5	99.5	905	647	654	264	129	191	0	6.5	437	432	776	3.3	242	212	90.8	131	635	311	310	328	432
London.St.Mary	797	758	561	718	189	190	93.5	56	5.5	309	108	829	154	443	806	637	643	669	160	322	265	745	182	17.4	95.9	862	614	650	259	129	186	6.5	0	431	433	770	4.1	240	209	95.5	131	623	307	305	330	390
Sunderland	465	381	533	386	324	464	366	439	437	289	453	801	320	24.2	746	264	252	238	291	612	197	297	395	427	343	442	183	263	353	507	458	437	431	0	726	391	444	504	238	523	404	227	306	236	123	379
Truro	1010	964	757	932	400	271	557	509	427	523	547	819	427	772	894	860	857	924	468	142	549	964	620	422	502	1127	866	864	430	320	279	432	433	726	0	974	464	319	493	514	592	838	524	519	628	567
Ulster	121	36.5	169	10.2	570	697	699	757	768	428	785	433	606	372	337	208	213	307	617	854	531	336	727	759	677	479	295	210	491	799	703	776	770	391	974	0	766	611	570	882	740	235	443	481	469	334
UCL	798	762	563	721	191	192	92.6	53.5	5.8	312	105	742	156	445	808	637	646	713	163	326	267	745	179	20.2	96.3	876	618	653	261	130	192	3.3	4.1	444	464	766	0	243	211	93.5	128	627	308	307	325	392
Cardiff	684	621	449	606	168	67.4	318	301	238	329	357	494	206	516	587	625	636	703	235	203	324	733	390	232	276	906	644	643	176	185	61.1	242	240	504	319	611	243	0	266	343	364	616	221	297	386	288
Nottingham	641	565	405	555	84.8	225	152	224	213	154	244	673	89.9	258	638	442	452	473	50.4	376	64	550	193	200	129	676	418	457	147	289	225	212	209	238	493	570	211	266	0	307	189	431	157	110	130	252
Ashford	920	867	690	504	284	280	165	106	93.3	446	153	958	261	535	918	769	752	778	254	404	360	836	252	145	183	981	694	759	369	199	278	90.8	95.5	523	514	882	93.5	343	307	0	189	732	430	414	415	499
W.Suffolk	808	726	576	715	204	314	47.7	72.8	140	318	47.8	844	172	415	811	612	609	628	163	468	242	720	66	136	68.1	837	578	617	275	251	314	131	131	404	592	740	128	364	189	189	0	591	312	297	299	405
Wishaw	306	222	386	229	441	576	551	623	620	361	636	650	476	207	588	27.2	25.8	83.8	485	721	383	135	579	610	528	278	71.7	32.6	425	666	569	635	623	227	838	235	627	616	431	732	591	0	383	334	321	452
Wrexham	490	448	263	420	111	239	284	335	310	22.1	370	516	148	321	470	402	402	470	171	391	145	511	359	299	249	654	411	408	49.9	336	225	311	307	306	524	443	308	221	157	430	312	383	0	383	199	100
Manchester	537	469	302	459	123	257	259	345	309	58.3	359	570	156	247	518	353	353	421	160	402	72.1	461	301	297	251	624	362	360	107	346	251	310	305	236	519	481	307	297	110	414	297	334	383	0	132	149
York	543	458	444	461	214	362	263	322	334	187	349	689	207	139	643	341	340	376	191	508	95	427	273	323	241	579	324	346	238	414	356	328	330	123	628	469	325	386	130	415	299	321	199	132	0	272
Bangor	392	342	157	314	208	342	367	431	420	101	482	425	242	391	369	476	472	538	265	448	218	584	447	419	382	742	480	478	132	431	348	432	390	379	567	334	392	288	252	499	405	452	100	149	272	0

Hospital	Truro	N.Devon	Bristol.Southmead	Bristol.RI	Southampton	London.Northwick	Chelsea	London.St.Mary	Ashford	London.St.Barts	UCL	Chelmsford	Colchester	Norfolk	W.Suffolk	Cambridge	Papworth	Leicester	Nottingham	Coventry	Birmingham	Shrewsbury	Chester	Manchester	Sheffield	York	Sunderland	Newcastle	Edinburgh.RI	Kirkcaldy	Dundee	Inverness	Glasgow.S.Gen	Glasgow.RI	Glasgow.Victoria	Wishaw	Altnaegelvin	Antrim	Belfast	Ulster	Dublin	Galway	Cork	Bangor	Wrexham	Cardiff
Truro	0	0.08	0.56	0.34	0.29	0.65	0.28	0.26	0.14	0.82	0.13	0.3	0.13	0.17	0.19	0.15	0.15	0.16	0.3	0.08	0.31	0.14	0.18	0.14	0.31	0.69	0.23	0.15	0.13	0.22	1	0.57	0.39	0.36	0.38	0.18	0.38	0.18	0.31	NA	0.25	0.24	0.27	0.27	0.09	0.44
N.Devon	0.08	0	0.63	0.34	0.24	0.69	0.38	0.35	0.15	0.84	0.22	0.29	0.14	0.12	0.11	0.08	0.08	0.11	0.19	0.04	0.27	0.09	0.15	0.09	0.33	0.66	0.15	0.07	0.08	0.19	1	0.51	0.41	0.37	0.39	0.12	0.38	0.18	0.32	NA	0.26	0.26	0.3	0.22	0	0.34
Bristol.Southmead	0.56	0.63	0	-0.09	0.2	0.6	0.25	0.18	-0.02	0.8	0.04	0.17	0.02	0.04	0.08	0.06	0.06	0.02	0.21	-0.06	0.17	0.01	0.02	-0.01	0.15	0.64	0.16	0.06	0	0.06	1	0.52	0.33	0.23	0.25	0.04	0.26	0.05	0.17	NA	0.09	0.09	0.13	0.18	-0.02	0.41
Bristol.RI	0.34	0.34	-0.09	0	0.1	0.57	0.21	0.11	-0.13	0.79	-0.03	0.08	-0.07	-0.08	-0.04	-0.06	-0.05	-0.09	0.09	-0.18	0.03	-0.1	-0.08	-0.14	0.06	0.6	0.07	-0.04	-0.1	-0.04	1	0.46	0.27	0.15	0.2	-0.06	0.18	-0.06	0.08	NA	0.02	0.01	0.06	0.07	-0.14	0.33
Southampton	0.29	0.24	0.2	0.1	0	0.66	0.34	0.28	0.09	0.83	0.15	0.22	0.12	0.07	0.08	0.07	0.06	0.04	0.25	-0.01	0.23	0.06	0.11	-0	0.26	0.64	0.15	0.04	0.05	0.15	1	0.51	0.38	0.31	0.33	0.09	0.34	0.12	0.27	NA	0.2	0.2	0.23	0.18	-0.02	0.4
London.Northwick	0.65	0.69	0.6	0.57	0.66	0	-0.1	-0.14	-0.06	0.71	-0.22	0.16	-0.13	0.01	0.06	0.03	0.03	0.01	0.12	-0.09	0.13	-0	0.01	-0.04	0.14	0.63	0.15	0.04	-0.01	0.06	1	0.51	0.26	0.22	0.25	0.03	0.25	0.02	0.15	NA	0.08	0.07	0.1	0.16	-0.04	0.4
Chelsea	0.28	0.38	0.25	0.21	0.34	-0.1	0	0.04	0.21	0.68	0	0.38	0.08	0.27	0.28	0.23	0.24	0.24	0.34	0.16	0.38	0.22	0.25	0.19	0.38	0.71	0.32	0.24	0.21	0.31	1	0.62	0.4	0.41	0.4	0.24	0.41	0.22	0.37	NA	0.3	0.28	0.32	0.34	0.2	0.56
London.St.Mary	0.26	0.35	0.18	0.11	0.28	-0.14	0.04	0	0.05	0.72	-0.08	0.23	-0.06	0.09	0.12	0.1	0.1	0.08	0.2	-0.01	0.23	0.08	0.09	0.03	0.23	0.66	0.2	0.1	0.06	0.15	1	0.55	0.31	0.28	0.29	0.12	0.32	0.09	0.22	NA	0.15	0.14	0.19	0.23	0.04	0.46
Ashford	0.14	0.15	-0.02	-0.13	0.09	-0.06	0.21	0.05	0	0.83	0.16	0.21	0.16	0.1	0.15	0.13	0.12	0.13	0.29	0.06	0.26	0.1	0.14	0.07	0.25	0.67	0.22	0.12	0.1	0.2	1	0.56	0.4	0.33	0.34	0.13	0.35	0.14	0.27	NA	0.21	0.2	0.18	0.25	0.07	0.48
London.St.Barts	0.82	0.84	0.8	0.79	0.83	0.71	0.68	0.72	0.83	0	-0.05	0.37	0.13	0.27	0.27	0.22	0.24	0.23	0.35	0.16	0.38	0.23	0.24	0.2	0.37	0.72	0.32	0.24	0.2	0.3	1	0.62	0.34	0.41	0.4	0.25	0.43	0.21	0.36	NA	0.29	0.27	0.31	0.35	0.2	0.55
UCL	0.13	0.22	0.04	-0.03	0.15	-0.22	0	-0.08	0.16	-0.05	0	0.24	-0.01	0.12	0.14	0.11	0.11	0.11	0.24	0.02	0.26	0.11	0.12	0.06	0.24	0.67	0.21	0.11	0.07	0.16	1	0.55	0.33	0.29	0.3	0.12	0.33	0.1	0.23	NA	0.15	0.15	0.2	0.23	0.06	0.47
Chelmsford	0.3	0.29	0.17	0.08	0.22	0.16	0.38	0.23	0.21	0.37	0.24	0	0.03	-0.2	0.05	0.02	0	0	0.13	-0.07	0.15	-0	0.02	-0.04	0.16	0.62	0.11	0.01	-0	0.08	1	0.49	0.34	0.25	0.27	0.01	0.28	0.06	0.19	NA	0.11	0.12	0.14	0.16	-0.11	0.39
Colchester	0.13	0.14	0.02	-0.07	0.12	-0.13	0.08	-0.06	0.16	0.13	-0.01	0.03	0	0.03	0.04	0.01	0	0.04	0.15	-0.05	0.18	0.01	0.07	0.02	0.21	0.64	0.12	0.04	0	0.11	1	0.49	0.35	0.27	0.3	0.07	0.29	0.06	0.21	NA	0.14	0.15	0.19	0.15	-0.04	0.38
Norfolk	0.17	0.12	0.04	-0.08	0.07	0.01	0.27	0.09	0.1	0.27	0.12	-0.2	0.03	0	0	-0.03	-0.05	-0.03	0.1	-0.09	0.11	-0.04	-0	-0.07	0.15	0.61	0.07	-0.04	-0.04	0.04	1	0.46	0.33	0.23	0.27	-0.01	0.25	0.03	0.17	NA	0.1	0.12	0.14	0.12	-0.18	0.35
W.Suffolk	0.19	0.11	0.08	-0.04	0.08	0.06	0.28	0.12	0.15	0.27	0.14	0.05	0.04	0	0	0.01	0.05	0.14	0.31	0.07	0.3	0.1	0.19	0.1	0.33	0.68	0.19	0.13	0.12	0.22	1	0.54	0.44	0.37	0.38	0.15	0.39	0.18	0.33	NA	0.27	0.26	0.29	0.19	0.07	0.47
Cambridge	0.15	0.08	0.06	-0.06	0.07	0.03	0.23	0.1	0.13	0.22	0.11	0.02	0.01	-0.03	0.01	0	0.08	0.2	0.37	0.14	0.36	0.14	0.26	0.18	0.39	0.7	0.24	0.17	0.17	0.28	1	0.57	0.46	0.41	0.43	0.21	0.42	0.23	0.38	NA	0.3	0.31	0.34	0.18	0.13	0.49
Papworth	0.15	0.08	0.06	-0.05	0.06	0.03	0.24	0.1	0.12	0.24	0.11	0	0	-0.05	0.05	0.08	0	0.14	0.32	0.06	0.31	0.09	0.2	0.11	0.34	0.68	0.19	0.11	0.12	0.22	1	0.55	0.43	0.37	0.39	0.16	0.39	0.18	0.33	NA	0.27	0.27	0.29	0.19	0.07	0.47
Leicester	0.16	0.11	0.02	-0.09	0.04	0.01	0.24	0.08	0.13	0.23	0.11	0	0.04	-0.03	0.14	0.2	0.14	0	0.26	-0.04	0.21	0.06	0.1	0.04	0.26	0.65	0.13	0.04	0.04	0.13	1	0.52	0.37	0.27	0.29	0.06	0.3	0.05	0.23	NA	0.18	0.15	0.22	0.19	-0	0.42
Nottingham	0.3	0.19	0.21	0.09	0.25	0.12	0.34	0.2	0.29	0.35	0.24	0.13	0.15	0.1	0.31	0.37	0.32	0.26	0	-0.21	-0.01	-0.2	-0.09	-0.15	0.08	0.56	-0.03	-0.14	-0.16	-0.07	1	0.39	0.28	0.15	0.21	-0.13	0.18	-0.06	0.1	NA	0.02	0.04	0.06	0	-0.3	0.17
Coventry	0.08	0.04	-0.06	-0.18	-0.01	-0.09	0.16	-0.01	0.06	0.16	0.02	-0.07	-0.05	-0.09	0.07	0.14	0.06	-0.04	-0.21	0	0.3	0.14	0.19	0.12	0.33	0.68	0.18	0.13	0.11	0.21	1	0.57	0.4	0.34	0.35	0.14	0.36	0.14	0.29	NA	0.22	0.21	0.26	0.26	0.1	0.5
Birmingham	0.31	0.27	0.17	0.03	0.23	0.13	0.38	0.23	0.26	0.38	0.26	0.15	0.18	0.11	0.3	0.36	0.31	0.21	-0.01	0.3	0	-0.1	-0.06	-0.12	0.1	0.59	0.03	-0.09	-0.09	-0.02	1	0.44	0.3	0.18	0.23	-0.07	0.21	-0.04	0.12	NA	0.06	0.04	0.09	0.08	-0.17	0.31
Shrewsbury	0.14	0.09	0.01	-0.1	0.06	-0	0.22	0.08	0.1	0.23	0.11	-0	0.01	-0.04	0.1	0.14	0.09	0.06	-0.2	0.14	-0.1	0	0.11	0.02	0.26	0.65	0.11	0.04	0.04	0.15	1	0.5	0.37	0.31	0.33	0.08	0.33	0.11	0.26	NA	0.18	0.19	0.24	0.11	-0.05	0.41
Chester	0.18	0.15	0.02	-0.08	0.11	0.01	0.25	0.09	0.14	0.24	0.12	0.02	0.07	-0	0.19	0.26	0.2	0.1	-0.09	0.19	-0.06	0.11	0	0.02	0.22	0.66	0.17	0.08	0.05	0.14	1	0.54	0.36	0.28	0.31	0.09	0.31	0.1	0.24	NA	0.16	0.16	0.21	0.21	-0.04	0.45
Manchester	0.14	0.09	-0.01	-0.14	-0	-0.04	0.19	0.03	0.07	0.2	0.06	-0.04	0.02	-0.07	0.1	0.18	0.11	0.04	-0.15	0.12	-0.12	0.02	0.02	0	0.35	0.68	0.23	0.15	0.15	0.25	1	0.58	0.42	0.38	0.38	0.16	0.39	0.2	0.33	NA	0.26	0.26	0.3	0.27	0.1	0.5
Sheffield	0.31	0.33	0.15	0.06	0.26	0.14	0.38	0.23	0.25	0.37	0.24	0.16	0.21	0.15	0.33	0.39	0.34	0.26	0.08	0.33	0.1	0.26	0.22	0.35	0	0.6	0.09	-0.01	-0.06	-0.01	1	0.48	0.3	0.17	0.23	-0.01	0.2	-0.02	0.11	NA	0.04	0.04	0.06	0.13	-0.1	0.36
York	0.69	0.66	0.64	0.6	0.64	0.63	0.71	0.66	0.67	0.72	0.67	0.62	0.64	0.61	0.68	0.7	0.68	0.65	0.56	0.68	0.59	0.65	0.66	0.68	0.6	0	0.2	0.12	0.13	0.23	1	0.56	0.43	0.38	0.38	0.16	0.38	0.18	0.33	NA	0.26	0.24	0.28	0.25	0.06	0.48
Sunderland	0.23	0.15	0.16	0.07	0.15	0.15	0.32	0.2	0.22	0.32	0.21	0.11	0.12	0.07	0.19	0.24	0.19	0.13	-0.03	0.18	0.03	0.11	0.17	0.23	0.09	0.2	0	0.03	0.13	0.24	1	0.54	0.45	0.41	0.42	0.17	0.41	0.18	0.36	NA	0.32	0.3	0.33	0.26	0.08	0.47
Newcastle	0.15	0.07	0.06	-0.04	0.04	0.04	0.24	0.1	0.12	0.24	0.11	0.01	0.04	-0.04	0.13	0.17	0.11	0.04	-0.14	0.13	-0.09	0.04	0.08	0.15	-0.01	0.12	0.03	0	0.15	0.25	1	0.56	0.44	0.39	0.4	0.18	0.4	0.18	0.35	NA	0.29	0.28	0.33	0.26	0.1	0.49
Edinburgh.RI	0.13	0.08	0	-0.1	0.05	-0.01	0.21	0.06	0.1	0.2	0.07	-0	0	-0.04	0.12	0.17	0.12	0.04	-0.16	0.11	-0.09	0.04	0.05	0.15	-0.06	0.13	0.13	0.15	0	0.18	1	0.45	0.36	0.32	0.29	0.16	0.37	0.14	0.28	NA	0.25	0.23	0.29	0.29	0.08	0.5
Kirkcaldy	0.22	0.19	0.06	-0.04	0.15	0.06	0.31	0.15	0.2	0.3	0.16	0.08	0.11	0.04	0.22	0.28	0.22	0.13	-0.07	0.21	-0.02	0.15	0.14	0.25	-0.01	0.23	0.24	0.25	0.18	0	1	0.48	0.29	0.16	0.2	0.01	0.24	-0.01	0.13	NA	0.1	0.08	0.16	0.15	-0.06	0.4
Dundee	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	NA	1	1	1	1	1	1
Inverness	0.57	0.51	0.52	0.46	0.51	0.51	0.62	0.55	0.56	0.62	0.55	0.49	0.49	0.46	0.54	0.57	0.55	0.52	0.39	0.57	0.44	0.5	0.54	0.58	0.48	0.56	0.54	0.56	0.45	0.48	1	0	0.41	0.33	0.36	0.06	0.36	0.12	0.29	NA	0.22	0.22	0.26	0.17	-0.13	0.38
Glasgow.S.Gen	0.39	0.41	0.33	0.27	0.38	0.26	0.4	0.31	0.4	0.34	0.33	0.34	0.35	0.33	0.44	0.46	0.43	0.37	0.28	0.4	0.3	0.37	0.36	0.42	0.3	0.43	0.45	0.44	0.36	0.29	1	0.41	0	0.34	-0.07	0.32	0.46	0.22	0.36	NA	0.34	0.31	0.38	0.43	0.31	0.63
Glasgow.RI	0.36	0.37	0.23	0.15	0.31	0.22	0.41	0.28	0.33	0.41	0.29	0.25	0.27	0.23	0.37	0.41	0.37	0.27	0.15	0.34	0.18	0.31	0.28	0.38	0.17	0.38	0.41	0.39	0.32	0.16	1	0.33	0.34	0	0.04	0.02	0.23	-0.07	0.09	NA	0.05	0.01	0.16	0.21	0.05	0.45
Glasgow.Victoria	0.38	0.39	0.25	0.2	0.33	0.25	0.4	0.29	0.34	0.4	0.3	0.27	0.3	0.27	0.38	0.43	0.39	0.29	0.21	0.35	0.23	0.33	0.31	0.38	0.23	0.38	0.42	0.4	0.29	0.2	1	0.36	-0.07	0.04	0	0.22	0.37	0.09	0.23	NA	0.23	0.18	0.28	0.36	0.22	0.57
Wishaw	0.18	0.12	0.04	-0.06	0.09	0.03	0.24	0.12	0.13	0.25	0.12	0.01	0.07	-0.01	0.15	0.21	0.16	0.06	-0.13	0.14	-0.07	0.08	0.09	0.16	-0.01	0.16	0.17	0.18	0.16	0.01	1	0.06	0.32	0.02	0.22	0	0.34	0.12	0.27	NA	0.23	0.2	0.26	0.24	0.06	0.47
Altnaegelvin	0.38	0.38	0.26	0.18	0.34	0.25	0.41	0.32	0.35	0.43	0.33	0.28	0.29	0.25	0.39	0.42	0.39	0.3	0.18	0.36	0.21	0.33	0.31	0.39	0.2	0.38	0.41	0.4	0.37	0.24	1	0.36	0.46	0.23	0.37	0.34	0	-0.12	0.03	NA	0.03	-0.04	0.13	0.26	0.09	0.49
Antrim	0.18	0.18	0.05	-0.06	0.12	0.02	0.22	0.09	0.14	0.21	0.1	0.06	0.06	0.03	0.18	0.23	0.18	0.05	-0.06	0.14	-0.04	0.11	0.1	0.2	-0.02	0.18	0.18	0.18	0.14	-0.01	1	0.12	0.22	-0.07	0.09	0.12	-0.12	0	0.17	NA	0.18	0.14	0.26	0.33	0.18	0.54
Belfast	0.31	0.32	0.17	0.08	0.27	0.15	0.37	0.22	0.27	0.36	0.23	0.19	0.21	0.17	0.33	0.38	0.33	0.23	0.1	0.29	0.12	0.26	0.24	0.33	0.11	0.33	0.36	0.35	0.28	0.13	1	0.29	0.36	0.09	0.23	0.27	0.03	0.17	0	NA	0.01	-0.04	0.12	0.23	0.06	0.47
Ulster	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Dublin	0.25	0.26	0.09	0.02	0.2	0.08	0.3	0.15	0.21	0.29	0.15	0.11	0.14	0.1	0.27	0.3	0.27	0.18	0.02	0.22	0.06	0.18	0.16	0.26	0.04	0.26	0.32	0.29	0.25	0.1	1	0.22	0.34	0.05	0.23	0.23	0.03	0.18	0.01	NA	0	-0	0.1	0.27	0.1	0.5
Galway	0.24	0.26	0.09	0.01	0.2	0.07	0.28	0.14	0.2	0.27	0.15	0.12	0.15	0.12	0.26	0.31	0.27	0.15	0.04	0.21	0.04	0.19	0.16	0.26	0.04	0.24	0.3	0.28	0.23	0.08	1	0.22	0.31	0.01	0.18	0.2	-0.04	0.14	-0.04	NA	-0	0	0.18	0.32	0.17	0.53
Cork	0.27	0.3	0.13	0.06	0.23	0.1	0.32	0.19	0.18	0.31	0.2	0.14	0.19	0.14	0.29	0.34	0.29	0.22	0.06	0.26	0.09	0.24	0.21	0.3	0.06	0.28	0.33	0.33	0.29	0.16	1	0.26	0.38	0.16	0.28	0.26	0.13	0.26	0.12	NA	0.1	0.18	0	0.3	0.14	0.52
Bangor	0.27	0.22	0.18	0.07	0.18	0.16	0.34	0.23	0.25	0.35	0.23	0.16	0.15	0.12	0.19	0.18	0.19	0.19	0	0.26	0.08	0.11	0.21	0.27	0.13	0.25	0.26	0.26	0.29	0.15	1	0.17	0.43	0.21	0.36	0.24	0.26	0.33	0.23	NA	0.27	0.32	0.3	0	0.04	0.45
Wrexham	0.09	0	-0.02	-0.14	-0.02	-0.04	0.2	0.04	0.07	0.2	0.06	-0.11	-0.04	-0.18	0.07	0.13	0.07	-0	-0.3	0.1	-0.17	-0.05	-0.04	0.1	-0.1	0.06	0.08	0.1	0.08	-0.06	1	-0.13	0.31	0.05	0.22	0.06	0.09	0.18	0.06	NA	0.1	0.17	0.14	0.04	0	0.42
Cardiff	0.44	0.34	0.41	0.33	0.4	0.4	0.56	0.46	0.48	0.55	0.47	0.39	0.38	0.35	0.47	0.49	0.47	0.42	0.17	0.5	0.31	0.41	0.45	0.5	0.36	0.48	0.47	0.49	0.5	0.4	1	0.38	0.63	0.45	0.57	0.47	0.49	0.54	0.47	NA	0.5	0.53	0.52	0.45	0.42	0

Supplementary Table A4. The Fixation Index (F_{ST}) is calculated for each pairwise MRSA hospital sub-populations. F_{ST} is a measure of population structure which compares the average number of pairwise differences between two sub-populations with the average number within the same sub-population. Values close to 0 indicate that the two populations are highly genetically similar, while values close to 1 imply that the two sub-populations are genetically distinct. It is possible to obtain negative values, which implies that the individuals from different sub-populations are genetically more similar than within sub-population.

	Truro	N.Devon	Bristol.Southmead	Bristol.RI	Southampton	London.Northwick	Chelsea	London.St.Mary	Ashford	London.St.Barts	UCL	Chelmsford	Colchester	Norfolk	W.Suffolk	Cambridge	Papworth	Leicester	Nottingham	Coventry	Birmingham	Shrewsbury	Chester	Manchester	Sheffield	York	Sunderland	Newcastle	Edinburgh.RI	Kirkcaldy	Dundee	Inverness	Glasgow.S.Gen	Glasgow.RI	Glasgow.Victoria	Wishaw	Altnaegelvin	Antrim	Belfast	Ulster	Dublin	Galway	Cork	Bangor	Wrexham	Cardiff
Truro	0.00	0.04	0.12	0.08	0.02	0.10	0.12	0.10	0.04	0.10	0.09	0.03	0.08	0.03	0.02	0.02	0.02	0.03	0.05	0.04	0.03	0.03	0.04	0.02	0.05	0.00	0.02	0.02	0.03	0.04	0.02	0.02	0.04	0.05	0.04	0.02	0.04	0.04	0.05	0.04	0.05	0.04	0.04	0.02	0.01	0.08
N.Devon	0.04	0.00	0.00	0.03	0.02	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.03	0.02	0.03	0.04	0.03	0.02	0.10	0.02	0.02	0.03	0.03	0.00	0.01	0.01	0.05	0.03	0.03	0.02	0.05	0.05	0.00	0.01	0.00	0.02	0.01	0.01	0.01	0.03	0.00	0.00	0.00	0.04	0.03	0.14
Bristol.Southmead	0.12	0.00	0.00	0.26	0.02	0.09	0.08	0.09	0.07	0.08	0.08	0.04	0.05	0.03	0.02	0.02	0.02	0.04	0.02	0.05	0.05	0.04	0.06	0.03	0.09	0.00	0.00	0.01	0.03	0.06	0.00	0.00	0.06	0.08	0.07	0.03	0.07	0.06	0.08	0.04	0.08	0.07	0.07	0.02	0.00	0.02
Bristol.RI	0.08	0.03	0.26	0.00	0.02	0.09	0.07	0.08	0.06	0.07	0.07	0.04	0.06	0.04	0.02	0.02	0.02	0.04	0.04	0.06	0.06	0.04	0.06	0.04	0.09	0.00	0.01	0.01	0.04	0.06	0.01	0.01	0.06	0.08	0.07	0.03	0.07	0.06	0.08	0.05	0.08	0.07	0.07	0.02	0.01	0.03
Southampton	0.02	0.02	0.02	0.02	0.00	0.02	0.01	0.02	0.02	0.01	0.02	0.01	0.02	0.02	0.01	0.01	0.02	0.03	0.02	0.02	0.03	0.02	0.02	0.03	0.02	0.01	0.02	0.02	0.01	0.02	0.01	0.02	0.01	0.02	0.02	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.03
London.Northwick	0.10	0.00	0.09	0.09	0.02	0.00	0.28	0.25	0.08	0.26	0.19	0.04	0.12	0.03	0.02	0.02	0.02	0.04	0.07	0.05	0.06	0.03	0.06	0.04	0.09	0.00	0.00	0.01	0.03	0.05	0.00	0.00	0.09	0.08	0.07	0.03	0.07	0.06	0.08	0.05	0.08	0.08	0.08	0.02	0.00	0.01
Chelsea	0.12	0.00	0.08	0.07	0.01	0.28	0.00	0.31	0.05	0.41	0.23	0.03	0.17	0.03	0.01	0.02	0.02	0.03	0.08	0.04	0.04	0.02	0.05	0.02	0.07	0.02	0.00	0.01	0.03	0.04	0.00	0.00	0.09	0.06	0.06	0.02	0.05	0.05	0.06	0.04	0.07	0.06	0.06	0.01	0.00	0.01
London.St.Mary	0.10	0.01	0.09	0.08	0.02	0.25	0.31	0.00	0.06	0.27	0.18	0.04	0.13	0.03	0.02	0.02	0.02	0.04	0.09	0.05	0.05	0.03	0.06	0.03	0.08	0.00	0.00	0.01	0.03	0.04	0.00	0.00	0.08	0.07	0.07	0.03	0.06	0.05	0.07	0.05	0.08	0.07	0.07	0.02	0.00	0.01
Ashford	0.04	0.00	0.07	0.06	0.02	0.08	0.05	0.06	0.00	0.05	0.06	0.07	0.04	0.06	0.02	0.02	0.02	0.03	0.03	0.03	0.04	0.03	0.05	0.02	0.08	0.00	0.01	0.01	0.02	0.04	0.00	0.00	0.04	0.05	0.05	0.02	0.05	0.04	0.05	0.04	0.06	0.05	0.08	0.02	0.01	0.01
London.St.Barts	0.10	0.00	0.08	0.07	0.01	0.26	0.41	0.27	0.05	0.00	0.23	0.03	0.13	0.03	0.01	0.02	0.01	0.03	0.07	0.04	0.04	0.02	0.05	0.02	0.06	0.00	0.00	0.01	0.03	0.04	0.00	0.00	0.14	0.06	0.06	0.02	0.05	0.05	0.06	0.04	0.07	0.06	0.06	0.01	0.00	0.01
UCL	0.09	0.01	0.08	0.07	0.02	0.19	0.23	0.18	0.06	0.23	0.00	0.04	0.12	0.03	0.02	0.02	0.02	0.03	0.07	0.05	0.05	0.03	0.05	0.03	0.08	0.00	0.00	0.01	0.04	0.05	0.01	0.01	0.09	0.07	0.07	0.03	0.07	0.06	0.08	0.04	0.08	0.08	0.07	0.02	0.00	0.02
Chelmsford	0.03	0.00	0.04	0.04	0.01	0.04	0.03	0.04	0.07	0.03	0.04	0.00	0.02	0.10	0.01	0.01	0.02	0.02	0.03	0.02	0.03	0.02	0.03	0.02	0.06	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.03	0.03	0.03	0.01	0.03	0.02	0.04	0.02	0.04	0.03	0.05	0.01	0.01	0.01
Colchester	0.08	0.03	0.05	0.06	0.02	0.12	0.17	0.13	0.04	0.13	0.12	0.02	0.00	0.04	0.04	0.05	0.04	0.03	0.07	0.04	0.04	0.04	0.04	0.02	0.05	0.01	0.02	0.02	0.03	0.04	0.03	0.03	0.04	0.04	0.04	0.02	0.04	0.04	0.05	0.05	0.05	0.04	0.04	0.04	0.02	0.04
Norfolk	0.03	0.02	0.03	0.04	0.02	0.03	0.03	0.03	0.06	0.03	0.03	0.10	0.04	0.00	0.02	0.02	0.03	0.02	0.04	0.03	0.03	0.03	0.03	0.02	0.05	0.00	0.02	0.01	0.02	0.03	0.02	0.02	0.02	0.03	0.03	0.02	0.03	0.02	0.03	0.03	0.03	0.03	0.03	0.02	0.04	0.03
W.Suffolk	0.02	0.03	0.02	0.02	0.01	0.02	0.01	0.02	0.02	0.01	0.02	0.01	0.04	0.02	0.00	0.09	0.07	0.02	0.02	0.03	0.02	0.05	0.02	0.01	0.02	0.00	0.03	0.02	0.02	0.03	0.03	0.03	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.04	0.02	0.02	0.02	0.07	0.01	0.03
Cambridge	0.02	0.04	0.02	0.02	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.05	0.02	0.09	0.00	0.10	0.03	0.02	0.03	0.02	0.06	0.01	0.01	0.02	0.00	0.04	0.02	0.02	0.03	0.04	0.04	0.01	0.02	0.01	0.02	0.02	0.02	0.02	0.05	0.02	0.01	0.02	0.10	0.02	0.04
Papworth	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.02	0.04	0.03	0.07	0.10	0.00	0.03	0.02	0.03	0.02	0.05	0.02	0.01	0.02	0.00	0.04	0.03	0.02	0.03	0.03	0.03	0.01	0.03	0.02	0.02	0.03	0.02	0.02	0.05	0.02	0.01	0.02	0.07	0.02	0.04
Leicester	0.03	0.02	0.04	0.04	0.03	0.04	0.03	0.04	0.03	0.03	0.03	0.02	0.03	0.02	0.02	0.03	0.03	0.00	0.02	0.05	0.04	0.04	0.03	0.02	0.04	0.01	0.04	0.04	0.03	0.04	0.02	0.02	0.04	0.06	0.05	0.03	0.05	0.06	0.06	0.04	0.05	0.05	0.04	0.03	0.01	0.03
Nottingham	0.05	0.10	0.02	0.04	0.02	0.07	0.08	0.09	0.03	0.07	0.07	0.03	0.07	0.04	0.02	0.02	0.02	0.02	0.00	0.02	0.02	0.03	0.02	0.01	0.03	0.01	0.02	0.02	0.02	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.02	0.02	0.02	0.03	0.07
Coventry	0.04	0.02	0.05	0.06	0.02	0.05	0.04	0.05	0.03	0.04	0.05	0.02	0.04	0.03	0.03	0.03	0.03	0.05	0.02	0.00	0.04	0.04	0.04	0.02	0.05	0.00	0.05	0.03	0.04	0.05	0.02	0.03	0.05	0.07	0.06	0.04	0.07	0.06	0.07	0.04	0.06	0.07	0.05	0.03	0.01	0.03
Birmingham	0.03	0.02	0.05	0.06	0.03	0.06	0.04	0.05	0.04	0.04	0.05	0.03	0.04	0.03	0.02	0.02	0.02	0.04	0.02	0.04	0.00	0.04	0.04	0.03	0.06	0.01	0.03	0.02	0.02	0.04	0.01	0.02	0.03	0.04	0.04	0.03	0.04	0.04	0.05	0.03	0.05	0.04	0.04	0.02	0.01	0.02
Shrewsbury	0.03	0.03	0.04	0.04	0.02	0.03	0.02	0.03	0.03	0.02	0.03	0.02	0.04	0.03	0.05	0.06	0.05	0.04	0.03	0.04	0.04	0.00	0.03	0.02	0.04	0.01	0.04	0.03	0.03	0.03	0.03	0.03	0.02	0.03	0.02	0.02	0.03	0.03	0.03	0.04	0.03	0.02	0.03	0.07	0.04	0.04
Chester	0.04	0.03	0.06	0.06	0.02	0.06	0.05	0.06	0.05	0.05	0.05	0.03	0.04	0.03	0.02	0.01	0.02	0.03	0.02	0.04	0.04	0.03	0.00	0.03	0.07	0.00	0.01	0.01	0.02	0.04	0.01	0.01	0.04	0.05	0.05	0.02	0.05	0.04	0.06	0.03	0.06	0.05	0.05	0.02	0.03	0.02
Manchester	0.02	0.00	0.03	0.04	0.03	0.04	0.02	0.03	0.02	0.02	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0.02	0.01	0.02	0.03	0.02	0.03	0.00	0.03	0.01	0.01	0.01	0.01	0.02	0.00	0.00	0.02	0.03	0.03	0.02	0.02	0.02	0.03	0.02	0.03	0.02	0.02	0.02	0.00	0.01
Sheffield	0.05	0.01	0.09	0.09	0.02	0.09	0.07	0.08	0.08	0.06	0.08	0.06	0.05	0.05	0.02	0.02	0.02	0.04	0.03	0.05	0.06	0.04	0.07	0.03	0.00	0.03	0.01	0.02	0.04	0.05	0.00	0.01	0.06	0.08	0.07	0.03	0.07	0.06	0.08	0.05	0.08	0.07	0.09	0.02	0.00	0.02
York	0.00	0.01	0.00	0.00	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.01	0.03	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
Sunderland	0.02	0.05	0.00	0.01	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.02	0.02	0.03	0.04	0.04	0.04	0.02	0.05	0.03	0.04	0.01	0.01	0.01	0.01	0.00	0.09	0.02	0.02	0.05	0.05	0.00	0.01	0.00	0.02	0.01	0.03	0.01	0.03	0.00	0.00	0.00	0.04	0.03	0.05
Newcastle	0.02	0.03	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.02	0.01	0.02	0.02	0.03	0.04	0.02	0.03	0.02	0.03	0.01	0.01	0.02	0.00	0.09	0.00	0.02	0.02	0.03	0.03	0.01	0.02	0.02	0.02	0.03	0.04	0.03	0.03	0.02	0.02	0.01	0.02	0.01	0.03
Edinburgh.RI	0.03	0.03	0.03	0.04	0.01	0.03	0.03	0.03	0.02	0.03	0.04	0.01	0.03	0.02	0.02	0.02	0.02	0.03	0.02	0.04	0.02	0.03	0.02	0.01	0.04	0.00	0.02	0.02	0.00	0.07	0.04	0.15	0.09	0.08	0.11	0.03	0.05	0.07	0.07	0.03	0.06	0.06	0.04	0.02	0.03	0.03
Kirkcaldy	0.04	0.02	0.06	0.06	0.02	0.05	0.04	0.04	0.04	0.04	0.05	0.02	0.04	0.03	0.03	0.03	0.03	0.04	0.03	0.05	0.04	0.03	0.04	0.02	0.05	0.00	0.02	0.02	0.07	0.00	0.06	0.03	0.09	0.11	0.11	0.04	0.08	0.08	0.10	0.04	0.08	0.07	0.06	0.03	0.01	0.03
Dundee	0.02	0.05	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.03	0.02	0.03	0.04	0.03	0.02	0.03	0.02	0.01	0.03	0.01	0.00	0.00	0.00	0.05	0.03	0.04	0.06	0.00	0.06	0.00	0.01	0.00	0.02	0.01	0.01	0.01	0.03	0.00	0.00	0.00	0.04	0.03	0.06
Inverness	0.02	0.05	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.03	0.02	0.03	0.04	0.03	0.02	0.03	0.03	0.02	0.03	0.01	0.00	0.01	0.00	0.05	0.03	0.15	0.03	0.06	0.00	0.00	0.01	0.00	0.03	0.01	0.01	0.01	0.03	0.00	0.00	0.00	0.04	0.07	0.06
Glasgow.S.Gen	0.04	0.00	0.06	0.06	0.01	0.09	0.09	0.08	0.04	0.14	0.09	0.03	0.04	0.02	0.01	0.01	0.01	0.04	0.02	0.05	0.03	0.02	0.04	0.02	0.06	0.00	0.00	0.01	0.09	0.09	0.00	0.00	0.00	0.19	0.40	0.04	0.09	0.11	0.13	0.04	0.10	0.11	0.08	0.01	0.00	0.01
Glasgow.RI	0.05	0.01	0.08	0.08	0.02	0.08	0.06	0.07	0.05	0.06	0.07	0.03	0.04	0.03	0.02	0.02	0.03	0.06	0.02	0.07	0.04	0.03	0.05	0.03	0.08	0.00	0.01	0.02	0.08	0.11	0.01	0.01	0.19	0.00	0.23	0.08	0.12	0.14	0.17	0.05	0.14	0.15	0.10	0.02	0.00	0.02
Glasgow.Victoria	0.04	0.00	0.07	0.07	0.02	0.07	0.06	0.07	0.05	0.06	0.07	0.03	0.04	0.03	0.02	0.01	0.02	0.05	0.02	0.06	0.04	0.02	0.05	0.03	0.07	0.00	0.00	0.02	0.11	0.11	0.00	0.00	0.40	0.23	0.00	0.05	0.12	0.14	0.16	0.05	0.13	0.14	0.09	0.01	0.00	0.01
Wishaw	0.02	0.02	0.03	0.03	0.01	0.03	0.02	0.03	0.02	0.02	0.03	0.01	0.02	0.02	0.02	0.02	0.02	0.03	0.02	0.04	0.03	0.02	0.02	0.03	0.03	0.00	0.02	0.02	0.03	0.04	0.02	0.03	0.04	0.08	0.05	0.00	0.05	0.05	0.05	0.03	0.05	0.05	0.04	0.02	0.01	0.02
Altnaegelvin	0.04	0.01	0.07	0.07	0.02	0.07	0.05	0.06	0.05	0.05	0.07	0.03	0.04	0.03	0.02	0.02	0.03	0.05	0.02	0.07	0.04	0.03	0.05	0.02	0.07	0.01	0.01	0.03	0.05	0.08	0.01	0.01	0.09	0.12	0.12	0.05	0.00	0.19	0.21	0.05	0.17	0.19	0.13	0.02	0.00	0.01
Antrim	0.04	0.01	0.06	0.06	0.02	0.06	0.05	0.05	0.04	0.05	0.06	0.02	0.04	0.02	0.02	0.02	0.02	0.06	0.02	0.06	0.04	0.03	0.04	0.02	0.06	0.01	0.03	0.04	0.07	0.08	0.01	0.01	0.11	0.14	0.14	0.05	0.19	0.00	0.19	0.05	0.13	0.14	0.10	0.02	0.01	0.02
Belfast	0.05	0.01	0.08	0.08	0.02	0.08	0.06	0.07	0.05	0.06	0.08	0.04	0.05	0.03	0.02	0.02	0.02	0.06	0.02	0.07	0.05	0.03	0.06	0.03	0.08	0.00	0.01	0.03	0.07	0.10	0.01	0.01	0.13	0.17	0.16	0.05	0.21	0.19	0.00	0.06	0.17	0.18	0.12	0.02	0.00	0.02
Ulster	0.04	0.03	0.04	0.05	0.02	0.05	0.04	0.05	0.04	0.04	0.04	0.02	0.05	0.03	0.04	0.05	0.05	0.04	0.03	0.04	0.03	0.04	0.03	0.02	0.05	0.00	0.03	0.03	0.03	0.04	0.03	0.03	0.04	0.05	0.05	0.03	0.05	0.05	0.06	0.00	0.05	0.05	0.05	0.05	0.01	0.04
Dublin	0.05	0.00	0.08	0.08	0.02	0.08	0.07	0.08	0.06	0.07	0.08	0.04	0.05	0.03	0.02	0.02	0.02	0.05	0.02	0.06	0.05	0.03	0.06	0.03	0.08	0.00	0.00	0.02	0.06	0.08	0.00	0.00	0.10	0.14	0.13	0.05	0.17	0.13	0.17	0.05	0.00	0.19	0.15	0.02	0.00	0.01
Galway	0.04	0.00	0.07	0.07	0.02	0.08	0.06	0.07	0.05	0.06	0.08	0.03	0.04	0.03	0.02	0.01	0.01	0.05	0.02	0.07	0.04	0.02	0.05	0.02	0.07	0.01	0.00	0.02	0.06	0.07	0.00	0.00	0.11	0.15	0.14	0.05	0.19	0.14	0.18	0.05	0.19	0.00	0.15	0.02	0.00	0.01
Cork	0.04	0.00	0.07	0.07	0.02	0.08	0.06	0.07	0.08	0.06	0.07	0.05	0.04	0.03	0.02	0.02	0.02	0.04	0.02	0.05	0.04	0.03	0.05	0.02	0.09	0.00	0.00	0.01	0.04	0.06	0.00	0.00	0.08	0.10	0.09	0.04	0.13	0.10	0.12	0.05	0.15	0.15	0.00	0.02	0.00	0.01
Bangor	0.02	0.04	0.02	0.02	0.02	0.02	0.01	0.02	0.02	0.01	0.02	0.01	0.04	0.02	0.07	0.10	0.07	0.03	0.02	0.03	0.02	0.07	0.02	0.02	0.02	0.00	0.04	0.02	0.02	0.03	0.04	0.04	0.01	0.02	0.01	0.02	0.02	0.02	0.02	0.02	0.05	0.02	0.02	0.02	0.00	0.02
Wrexham	0.01	0.03	0.00	0.01	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.02	0.04	0.01	0.02	0.02	0.01	0.03	0.01	0.01	0.04	0.03	0.00	0.00	0.00	0.03	0.01	0.03	0.01	0.03	0.07	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.02	0.00	0.03
Cardiff	0.08	0.14	0.02	0.03	0.03	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.04	0.03	0.03	0.04	0.04	0.03	0.07	0.03	0.02	0.04	0.02	0.01	0.02	0.00	0.05	0.03	0.03	0.03	0.06	0.06	0.01	0.02	0.01	0.02	0.01	0.02	0.02	0.04	0.01	0.01	0.01	0.04	0.03	0.00

Supplementary Table A5. The $SNP_{SIMILARITY}$ matrix. I compare the SNPs contained in one isolate with those of another isolate. Each SNP found in both isolates indicates increased similarity between the two isolates. The number of SNPs found in both isolates is standardised by the sum of the number of SNPs harboured in each isolate.

	Truro	N.Devon	Bristol.Southmead	Bristol.RI	Southampton	London.Northwick	Chelsea	London.St.Mary	Ashford	London.St.Barts	UCL	Chelmsford	Colchester	Norfolk	W.Suffolk	Cambridge	Papworth	Leicester	Nottingham	Coventry	Birmingham	Shrewsbury	Chester	Manchester	Sheffield	York	Sunderland	Newcastle	Edinburgh.RI	Kirkcaldy	Dundee	Inverness	Glasgow.S.Gen	Glasgow.RI	Glasgow.Victoria	Wishaw	Altnaegelvin	Antrim	Belfast	Ulster	Dublin	Galway	Cork	Bangor	Wrexham	Cardiff
Truro	327	30	20	68	90	25	35	67	65	23	86	11	41	31	22	102	88	45	57	32	35	51	36	46	42	5	49	42	10	13	3	6	25	13	7	14	33	32	41	6	48	5	25	31	7	61
N.Devon	30	66	0	21	34	1	1	24	3	0	13	0	5	3	5	17	28	8	32	4	5	14	46	4	13	1	13	13	3	3	3	3	1	3	0	4	4	4	6	0	3	0	3	17	3	19
Bristol.Southmead	20	0	42	40	6	5	6	7	19	6	7	5	7	5	7	25	9	5	5	8	6	7	6	5	5	0	0	7	6	10	0	1	6	6	5	6	5	7	8	6	6	5	5	5	0	25
Bristol.RI	68	21	40	227	61	5	6	47	36	7	24	14	12	15	23	62	81	19	27	36	19	23	41	72	28	0	12	20	16	21	4	4	14	19	12	15	24	17	29	10	26	11	18	64	4	41
Southampton	90	34	6	61	410	6	6	73	23	6	62	13	18	15	33	119	135	104	43	77	55	60	61	92	76	25	49	52	11	20	3	9	10	16	11	11	50	29	43	8	24	13	24	72	3	100
London.Northwick	25	1	5	5	6	50	24	31	30	22	42	7	24	5	5	35	25	10	24	6	8	6	6	21	7	1	1	6	5	5	0	1	24	5	5	5	6	6	6	5	24	7	21	5	0	7
Chelsea	35	1	6	6	6	24	72	67	13	31	41	5	32	5	5	37	19	10	27	6	6	10	5	13	6	5	6	7	6	5	0	1	31	7	5	5	5	6	11	6	32	5	13	6	0	12
London.St.Mary	67	24	7	47	73	31	67	316	37	34	120	21	66	16	41	84	84	20	58	54	20	28	42	85	23	3	14	23	16	17	3	4	39	16	13	13	28	17	25	10	51	15	44	34	6	24
Ashford	65	3	19	36	23	30	13	37	352	13	64	38	28	33	47	153	129	28	27	36	25	37	29	52	25	2	47	22	18	25	3	7	22	25	13	23	30	17	31	9	68	9	71	33	6	34
London.St.Barts	23	0	6	7	6	22	31	34	13	53	41	5	23	5	5	42	16	5	22	5	6	7	5	13	5	0	1	6	6	5	0	1	51	5	5	5	5	6	9	6	23	5	13	5	0	6
UCL	86	13	7	24	62	42	41	120	64	41	369	19	72	15	73	135	120	41	43	79	40	86	34	62	47	5	19	40	34	47	3	6	58	34	23	15	55	38	52	22	99	41	79	41	5	55
Chelmsford	11	0	5	14	13	7	5	21	38	5	19	104	5	46	13	56	60	7	14	6	15	13	12	14	10	0	1	6	5	6	0	0	5	5	5	6	13	5	11	5	20	7	18	26	3	8
Colchester	41	5	7	12	18	24	32	66	28	23	72	5	134	13	17	64	38	11	34	24	11	20	16	18	17	3	9	14	10	12	3	5	23	16	6	8	19	9	20	6	56	7	25	19	3	17
Norfolk	31	3	5	15	15	5	5	16	33	5	15	46	13	75	13	54	34	10	17	11	14	13	22	16	11	0	7	10	8	9	3	3	6	11	6	9	14	8	14	5	13	5	12	17	10	12
W.Suffolk	22	5	7	23	33	5	5	41	47	5	73	13	17	13	443	244	208	55	9	87	31	71	37	39	24	4	69	17	9	19	3	5	8	26	11	18	43	14	33	5	45	10	27	73	3	26
Cambridge	102	17	25	62	119	35	37	84	153	42	135	56	64	54	244	1438	601	123	51	126	60	109	62	104	153	39	74	62	21	36	8	8	53	52	22	26	76	37	70	12	108	18	81	93	15	116
Papworth	88	28	9	81	135	25	19	84	129	16	120	60	38	34	208	601	1250	116	56	122	59	105	70	100	104	10	64	72	21	37	3	10	26	51	21	32	115	44	119	18	98	21	81	107	7	91
Leicester	45	8	5	19	104	10	10	20	28	5	41	7	11	10	55	123	116	368	14	71	52	55	53	63	75	6	58	44	15	21	3	7	13	19	14	25	47	34	45	13	40	16	27	41	3	96
Nottingham	57	32	5	27	43	24	27	58	27	22	43	14	34	17	9	51	56	14	89	11	12	21	34	20	25	2	16	18	8	8	3	4	22	9	5	8	10	9	10	5	33	7	17	19	6	27
Coventry	32	4	8	36	77	6	6	54	36	5	79	6	24	11	87	126	122	71	11	496	71	72	61	48	51	7	74	62	19	30	4	7	16	30	19	17	66	41	55	16	49	21	53	36	4	78
Birmingham	35	5	6	19	55	8	6	20	25	6	40	15	11	14	31	60	59	52	12	71	200	56	36	48	31	5	38	38	9	11	4	7	9	9	7	22	36	28	35	6	26	11	15	31	3	41
Shrewsbury	51	14	7	23	60	6	10	28	37	7	86	13	20	13	71	109	105	55	21	72	56	392	98	46	50	6	42	42	17	17	3	11	8	21	6	19	47	29	46	6	45	11	23	86	53	64
Chester	36	46	6	41	61	6	5	42	29	5	34	12	16	22	37	62	70	53	34	61	36	98	393	36	40	2	24	25	17	16	3	9	10	14	13	16	27	15	23	7	40	10	29	32	61	55
Manchester	46	4	5	72	92	21	13	85	52	13	62	14	18	16	39	104	100	63	20	48	48	46	36	408	75	19	47	72	17	21	4	8	20	18	13	22	43	35	40	11	53	11	38	85	8	36
Sheffield	42	13	5	28	76	7	6	23	25	5	47	10	17	11	24	153	104	75	25	51	31	50	40	75	359	32	44	57	32	79	3	7	24	29	31	20	51	48	52	13	35	15	29	44	6	40
York	5	1	0	0	25	1	5	3	2	0	5	0	3	0	4	39	10	6	2	7	5	6	2	19	32	46	13	3	0	5	0	1	1	0	1	0	4	3	6	0	2	2	5	4	0	9
Sunderland	49	13	0	12	49	1	6	14	47	1	19	1	9	7	69	74	64	58	16	74	38	42	24	47	44	13	370	129	11	15	5	8	5	12	4	9	31	27	30	2	21	2	15	29	9	43
Newcastle	42	13	7	20	52	6	7	23	22	6	40	6	14	10	17	62	72	44	18	62	38	42	25	72	57	3	129	355	19	42	3	10	17	27	34	24	61	62	64	35	47	21	40	37	5	33
Edinburgh.RI	10	3	6	16	11	5	6	16	18	6	34	5	10	8	9	21	21	15	8	19	9	17	17	17	32	0	11	19	129	73	5	29	26	30	27	15	15	28	29	12	46	11	15	9	9	20
Kirkcaldy	13	3	10	21	20	5	5	17	25	5	47	6	12	9	19	36	37	21	8	30	11	17	16	21	79	5	15	42	73	293	8	3	25	90	31	24	57	69	70	30	65	21	39	20	3	28
Dundee	3	3	0	4	3	0	0	3	3	0	3	0	3	3	3	8	3	3	3	4	4	3	3	4	3	0	5	3	5	8	26	3	1	5	0	3	3	3	4	0	6	1	4	3	3	4
Inverness	6	3	1	4	9	1	1	4	7	1	6	0	5	3	5	8	10	7	4	7	7	11	9	8	7	1	8	10	29	3	3	72	1	3	0	16	7	6	6	1	6	1	4	7	8	7
Glasgow.S.Gen	25	1	6	14	10	24	31	39	22	51	58	5	23	6	8	53	26	13	22	16	9	8	10	20	24	1	5	17	26	25	1	1	135	45	66	12	12	24	30	12	33	11	29	6	0	11
Glasgow.RI	13	3	6	19	16	5	7	16	25	5	34	5	16	11	26	52	51	19	9	30	9	21	14	18	29	0	12	27	30	90	5	3	45	174	48	25	27	52	62	14	36	11	27	25	3	20
Glasgow.Victoria	7	0	5	12	11	5	5	13	13	5	23	5	6	6	11	22	21	14	5	19	7	6	13	13	31	1	4	34	27	31	0	0	66	48	101	13	16	26	25	13	19	11	17	11	0	12
Wishaw	14	4	6	15	11	5	5	13	23	5	15	6	8	9	18	26	32	25	8	17	22	19	16	22	20	0	9	24	15	24	3	16	12	25	13	123	23	24	25	11	23	11	14	19	3	12
Altnaegelvin	33	4	5	24	50	6	5	28	30	5	55	13	19	14	43	76	115	47	10	66	36	47	27	43	51	4	31	61	15	57	3	7	12	27	16	23	264	60	110	31	62	54	47	40	4	26
Antrim	32	4	7	17	29	6	6	17	17	6	38	5	9	8	14	37	44	34	9	41	28	29	15	35	48	3	27	62	28	69	3	6	24	52	26	24	60	180	134	46	38	21	36	15	4	16
Belfast	41	6	8	29	43	6	11	25	31	9	52	11	20	14	33	70	119	45	10	55	35	46	23	40	52	6	30	64	29	70	4	6	30	62	25	25	110	134	302	45	57	23	47	34	4	30
Ulster	6	0	6	10	8	5	6	10	9	6	22	5	6	5	5	12	18	13	5	16	6	6	7	11	13	0	2	35	12	30	0	1	12	14	13	11	31	46	45	53	32	21	29	7	0	8
Dublin	48	3	6	26	24	24	32	51	68	23	99	20	56	13	45	108	98	40	33	49	26	45	40	53	35	2	21	47	46	65	6	6	33	36	19	23	62	38	57	32	395	37	160	36	4	33
Galway	5	0	5	11	13	7	5	15	9	5	41	7	7	5	10	18	21	16	7	21	11	11	10	11	15	2	2	21	11	21	1	1	11	11	11	11	54	21	23	21	37	101	62	13	0	12
Cork	25	3	5	18	24	21	13	44	71	13	79	18	25	12	27	81	81	27	17	53	15	23	29	38	29	5	15	40	15	39	4	4	29	27	17	14	47	36	47	29	160	62	324	30	3	25
Bangor	31	17	5	64	72	5	6	34	33	5	41	26	19	17	73	93	107	41	19	36	31	86	32	85	44	4	29	37	9	20	3	7	6	25	11	19	40	15	34	7	36	13	30	354	3	55
Wrexham	7	3	0	4	3	0	0	6	6	0	5	3	3	10	3	15	7	3	6	4	3	53	61	8	6	0	9	5	9	3	3	8	0	3	0	3	4	4	4	0	4	0	3	3	62	9
Cardiff	61	19	25	41	100	7	12	24	34	6	55	8	17	12	26	116	91	96	27	78	41	64	55	36	40	9	43	33	20	28	4	7	11	20	12	12	26	16	30	8	33	12	25	55	9	369

Supplementary Table A6. The $SNP_{CONNECTIVITY}$ matrix. For each pair of hospitals I determine which of the 5469 SNPs appear in both. This is now the SNP connectivity of those two hospitals. A higher number of SNPs appearing in both hospitals indicates a greater level of connectivity between the sub-populations. A number of the SNPs provide connectivity within a hospital by being harboured in multiple isolates in the same hospital.

	Truro	N.Devon	Bristol.Southmead	Bristol.RI	Southampton	London.Northwick	Chelsea	London.St.Mary	Ashford	London.St.Barts	UCL	Chelmsford	Colchester	Norfolk	W.Suffolk	Cambridge	Papworth	Leicester	Nottingham	Coventry	Birmingham	Shrewsbury	Chester	Manchester	Sheffield	York	Sunderland	Newcastle	Edinburgh.RI	Kirkcaldy	Dundee	Inverness	Glasgow.S.Gen	Glasgow.RI	Glasgow.Victoria	Wishaw	Altnaegelvin	Antrim	Belfast	Ulster	Dublin	Galway	Cork	Bangor	Wrexham	Cardiff
Truro	101	0	0	9	3	0	0	1	1	0	3	0	0	0	0	2	3	2	0	0	0	0	0	0	0	0	3	1	0	0	0	0	0	1	0	1	0	0	1	0	2	0	1	0	0	2
N.Devon	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Bristol.Southmead	0	0	13	12	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bristol.RI	9	0	12	78	0	0	0	1	0	0	0	0	0	0	1	14	23	0	0	0	0	0	2	2	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Southampton	3	0	0	0	121	0	0	23	0	0	0	0	1	0	2	3	0	8	0	5	0	0	1	0	3	0	1	0	1	0	0	1	0	1	0	1	0	0	0	3	0	0	0	0	0	7
London.Northwick	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Chelsea	0	0	0	0	0	0	10	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
London.St.Mary	1	0	0	1	23	0	10	98	0	0	2	0	23	0	0	10	0	0	0	0	0	2	0	21	0	0	0	0	0	0	0	0	0	1	0	0	2	0	0	0	0	1	0	0	0	0
Ashford	1	0	0	0	0	0	0	0	91	0	1	0	0	0	3	21	3	1	0	0	0	0	0	1	1	0	32	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0
London.St.Barts	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0
UCL	3	0	0	0	0	0	0	2	1	0	64	0	11	0	1	0	5	0	1	0	6	29	0	1	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1
Chelmsford	0	0	0	0	0	0	0	0	0	0	0	21	0	2	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Colchester	0	0	0	0	1	0	0	23	0	0	11	0	56	3	0	11	1	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Norfolk	0	0	0	0	0	0	0	0	0	0	0	2	3	8	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
W.Suffolk	0	0	0	1	2	0	0	0	3	0	1	0	0	0	78	25	17	1	0	0	0	6	1	1	0	0	4	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Cambridge	2	0	1	14	3	0	0	10	21	1	0	0	11	0	25	571	111	12	0	18	1	1	5	1	18	2	1	1	1	1	0	0	0	1	0	1	5	0	4	0	2	1	2	2	0	17
Papworth	3	0	0	23	0	0	0	0	3	0	5	9	1	0	17	111	383	0	1	9	0	0	2	1	0	1	4	0	0	0	0	1	0	0	0	2	1	0	25	0	2	0	0	3	0	5
Leicester	2	0	0	0	8	4	0	0	1	0	0	0	0	0	1	12	0	126	1	0	0	0	4	0	4	0	11	1	0	0	0	0	0	0	1	1	0	0	1	0	1	0	1	0	0	26
Nottingham	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
Coventry	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	18	9	0	0	135	4	0	1	0	1	1	0	1	0	1	0	0	0	1	1	0	1	2	1	0	0	0	1	0	2
Birmingham	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	1	0	0	0	4	71	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0
Shrewsbury	0	0	0	0	0	0	0	2	0	0	29	0	0	0	6	1	0	0	0	0	0	120	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	8
Chester	0	0	0	2	1	0	0	0	0	0	0	0	0	2	1	5	2	4	0	1	0	0	90	0	1	0	1	0	0	0	0	0	0	1	0	2	0	2	0	2	0	1	0	0	4
Manchester	0	0	0	2	0	0	0	21	1	0	1	0	0	0	0	1	1	0	0	0	0	0	0	83	1	0	1	12	0	0	0	0	0	0	0	0	0	0	4	0	0	1	0	3
Sheffield	0	0	0	0	3	0	0	0	1	0	0	0	5	0	0	18	0	4	0	1	0	1	1	1	110	2	0	0	0	37	0	0	0	0	0	2	0	0	0	0	0	1	1	1	0
York	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	1	0	0	0	0	2	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Sunderland	3	0	0	1	1	0	0	0	32	0	0	0	0	0	4	1	4	11	0	0	0	2	1	1	0	0	135	30	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	2
Newcastle	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	12	0	0	30	143	0	0	0	0	0	0	0	9	0	0	0	0	0	1	0	0	1
Edinburgh.RI	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	25	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Kirkcaldy	0	0	0	1	0	0	0	0	1	0	2	0	0	0	1	1	0	0	0	1	0	0	0	0	37	0	0	0	6	92	0	0	1	20	0	0	17	0	0	0	0	0	1	0	1
Dundee	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Inverness	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Glasgow.S.Gen	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	24	0	1	0	0	0	0	0	0	6	0	0	0	0
Glasgow.RI	1	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	20	0	0	0	44	0	0	0	0	0	0	0	0	0	0	0	0
Glasgow.Victoria	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	9	0	0	0	0	1	0	14	0	0	0	0	0	0	0	0	0	0	1
Wishaw	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	2	1	0	0	2	0	0	0	2	0	0	0	0	0	0	0	0	0	0	63	1	0	1	0	0	0	0	0	0	0
Altnaegelvin	0	0	0	0	0	0	0	2	0	0	0	0	1	0	0	5	1	0	0	1	0	0	2	0	0	0	1	0	0	17	0	0	0	0	0	0	1	85	0	10	0	0	24	0	0	0
Antrim	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	47	11	5	0	0	0	0	0
Belfast	1	0	0	0	3	0	0	0	0	0	0	0	0	0	0	4	25	1	0	1	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	1	10	11	92	0	0	1	1	1	0	2
Ulster	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	5	0	0	0	0	0	0
Dublin	2	0	0	0	0	0	0	0	1	0	0	1	0	0	0	2	2	1	1	0	0	2	2	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	108	1	18	0	0	0
Galway	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24	0	1	0	1	53	26	0	0	0
Cork	1	0	0	0	0	0	0	1	1	0	1	0	0	0	0	2	0	1	0	0	0	0	1	0	1	0	0	1	0	0	0	0	6	0	0	0	0	0	1	0	18	26	114	3	0	1
Bangor	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	3	0	0	1	0	0	0	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	3	86	0	10
Wrexham	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Cardiff	2	0	0	0	7	1	0	0	0	0	1	0	0	0	0	17	5	26	1	2	0	8	4	3	0	1	2	1	0	1	0	0	0	0	1	0	0	0	2	0	0	0	1	10	0	138

Supplementary Table A7. The majority of SNPs (n = 2655) are only ever seen in two isolates. These rare SNPs might be more informative in the SNP connectivity since I are removing those more common SNPs which might be present in all sub-populations, and therefore contribute "noise" to the network. For each pair of hospitals I determine which of these rare SNPs appear in both. This is now the Rare SNP connectivity of those two hospitals.

Supplementary Table A8. I extracted the relevant Patient Referral data from Donker *et al* (2012, 2014) of the 27 English hospitals which appear in this thesis where I also have genetic information.

	Bristol.RI	N.Devon	York	Colchester	Truro	W.Suffolk	Cambridge	Southampton	Sheffield	Chester	Coventry	Sunderland	Norfolk	Manchester	London.St.Barts	Chelmsford	Chelsea	UCL	Newcastle	Ashford	London.Northwick	Bristol.Southmead	Leicester	Nottingham	Birmingham	Shrewsbury	London.St.Mary
Bristol.RI	249994	562	5	7	837	6	33	90	27	9	33	4	22	3	24	11	3	47	19	12	28	23198	25	29	26	16	85
N.Devon	533	85401	4	6	153	2	14	31	8	1	21	3	8	3	14	0	8	23	3	14	14	221	30	19	8	14	25
York	14	5	134785	8	28	11	29	10	88	9	14	30	16	17	8	12	9	21	142	6	7	9	23	35	6	12	19
Colchester	7	9	6	136310	16	244	599	11	9	1	7	2	156	3	1234	5	2350	229	8	9	37	4	18	13	4	5	110
Truro	737	141	17	19	310174	10	40	88	62	16	44	4	20	19	55	17	26	154	14	33	27	301	73	87	29	39	90
W.Suffolk	5	8	5	233	14	97233	4506	7	16	4	10	4	622	5	48	2	100	46	4	7	22	7	16	23	3	6	32
Cambridge	41	7	19	731	45	5256	325666	84	58	19	33	9	1650	7	150	12	477	187	37	16	67	19	106	103	19	25	192
Southampton	95	32	11	10	80	10	85	259616	16	8	22	7	22	5	52	16	18	73	16	48	37	39	32	28	9	20	75
Sheffield	11	12	116	8	62	11	54	49	608013	42	49	21	66	107	20	64	4	43	143	8	13	37	213	456	72	54	32
Chester	4	6	8	6	18	4	22	7	47	167193	11	2	6	74	3	624	5	13	7	6	9	10	19	15	7	115	12
Coventry	39	16	6	9	65	13	27	27	45	19	214924	12	29	15	28	16	5	75	32	8	35	18	1011	108	310	55	48
Sunderland	5	1	22	6	3	3	8	12	23	4	4	207303	6	8	4	8	4	12	6138	1	8	1	11	16	10	4	14
Norfolk	18	9	20	178	25	720	1519	22	51	5	39	6	408353	4	119	8	104	141	17	21	45	16	80	76	9	8	93
Manchester	7	4	16	4	29	2	3	7	83	106	10	8	5	115453	6	76	3	24	39	2	11	7	29	21	5	36	23
London.St.Barts	45	11	3	1479	62	36	151	54	28	7	17	3	77	5	168389	10	1627	1189	21	39	200	21	51	26	29	6	542
Chelmsford	12	2	5	3	14	3	10	18	64	674	19	8	3	84	17	204787	4	22	26	3	5	6	17	23	17	36	13
Chelsea	9	10	5	2405	27	82	454	18	10	4	5	5	97	1	1361	5	144302	405	12	12	67	12	16	11	4	6	96
UCL	57	27	23	275	197	46	201	105	50	14	84	9	120	27	1341	20	480	153247	47	589	1402	58	136	73	24	29	2301
Newcastle	24	6	146	13	18	5	44	15	122	20	34	7369	17	27	22	34	5	38	310548	10	7	19	34	83	21	13	36
Ashford	21	17	4	9	40	14	19	43	14	8	9	4	17	3	43	6	11	482	6	88938	88	11	7	17	5	3	764
London.Northwick	29	10	4	33	25	17	58	35	19	3	33	7	27	15	214	9	55	1120	5	91	146872	26	69	23	24	11	7571
Bristol.Southmead	20556	250	8	8	383	6	20	53	13	7	19	10	14	8	25	8	9	54	21	8	25	376519	33	25	23	20	42
Leicester	28	24	28	16	71	21	108	40	192	19	1108	13	86	29	62	15	21	102	37	9	88	23	464062	2508	148	44	62
Nottingham	34	9	36	16	86	25	93	28	393	21	130	15	72	32	26	16	10	67	83	19	25	24	2623	379656	69	36	41
Birmingham	38	7	8	5	28	7	14	19	60	8	264	5	9	8	31	16	5	23	16	3	24	26	137	73	235250	142	39
Shrewsbury	21	10	14	10	47	3	21	21	56	106	56	1	12	35	8	38	5	30	13	9	13	14	40	36	144	208120	30
London.St.Mary	82	36	17	120	90	32	162	76	32	9	40	8	72	15	534	13	99	1966	28	677	8394	46	56	44	49	31	303207

Appendix B

Supplementary Table B1. There are 18 Candidate Introductions (CIs) at the hospital geographic resolution which also harbour a LSS for the TAPO posited origin location. 12 of the CIs are between Referral Clusters (RCs), while the rest are within RCs.

Strain Code	Year	Sampling Location	TAPO	Referral Cluster
8728 5.17	2001	Cork	Coventry	Between
7554 6.12	2001	Newcastle	Sunderland	Within
7922 1.38	2001	Papworth	Ashford	Between
7414 7.72	2002	West Suffolk	Cambridge	Within
8728 5.23	2003	Kirkcaldy	Altnaegelvin	Between
7712 8.69	2004	Bristol Royal Infirmary	Manchester	Between
7414 8.74	2004	Shrewsbury	Cardiff	Between
7521 5.13	2005	Altnaegelvin	Papworth	Between
7521 5.58	2006	Kirkcaldy	Glasgow Royal Infirmary	Within
7521 6.76	2007	Southampton	Truro	Between
7065 8.76	2007	Cambridge	Bristol Royal Infirmary	Between
7469 7.7	2008	Ashford	Cork	Between
7469 7.9	2008	Ashford	Truro	Between
7480 8.19	2008	UCL	Cork	Between
7469 7.15	2008	Bangor	Manchester	Between
7748 6.65	2009	West Suffolk	Papworth	Within
7083 1.16	2009	Cambridge	Papworth	Within
7564 8.93	2010	Bristol Southmead	Bristol	Within

Supplementary Table B2. There are 16 Candidate Introductions (CIs) at the Referral Cluster (RC) geographic resolution which also harbour a LSS for the TAPO posited origin location. 10 of the CIs are to RCs that are non-adjacent. The rest are between adjacent RCs or RCs which share a sea border.

Strain Code	Year	Sampling Location	TAPO	Referral Cluster
8728 5.17	2001	RC 13	RC 6	Non-adjacent
7922 1.38	2001	RC 8	RC 1	Non-adjacent
7712 8.38	2003	RC 4	RC 8	Adjacent
8728 5.23	2003	RC 15	RC 14	Adjacent via sea
7712 8.69	2004	RC 2	RC 12	Non-adjacent
7414 8.74	2004	RC 6	RC 16	Adjacent
7521 5.13	2005	RC 14	RC 8	Non-adjacent
7521 6.76	2007	RC 5	RC 2	Adjacent
7065 8.70	2007	RC 8	RC 15	Non-adjacent
7065 8.76	2007	RC 8	RC 2	Non-adjacent
7469 7.9	2008	RC 1	RC 2	Non-adjacent
7480 8.19	2008	RC 4	RC 13	Non-adjacent
7469 7.15	2008	RC 16	RC 12	Non-adjacent
7480 8.61	2009	RC 16	RC 2	Adjacent
7564 8.78	2004	RC 10	RC 14	Adjacent via sea
7564 8.68	2010	RC 15	RC 4	Non-adjacent

Appendix C

Supplementary Table C1. The summary of all 127 Candidate Introductions with the predictions for both TAPO and SnAPO. For the 41 isolates where there is contradiction between the two methods I provide an explanation why there is a discrepancy. In all 41 cases the discrepancy could be attributed to the TAPO method, with 22 potential CIs sampled before the other isolates in the sub-clade, while the 19 possible CIs could be grouped into different sub-clades.

Strain Code	Date Sampled	Hospital Sampled	TAPO (Hospital)	SnAPO (Hospital)	Reason Different?
X7414_7.33	08/01/01	Chelmsford	Papworth	Manchester	Sampled before all Papworth isolates in subclade.
X8728_5.17	08/01/01	Cork	Coventry	Cork	Sampled before all Coventry isolates in subclade.
X7554_6.39	13/01/01	Shrewsbury	Birmingham	Southampton	Sampled before all Birmingham isolates in subclade.
X7554_6.29	29/01/01	Norfolk	Truro	Southampton	Sampled before all Truro isolates in subclade.
X7414_7.34	24/02/01	Chelmsford	Papworth	Chelmsford	Sampled before all Papworth isolates in subclade.
X7554_6.12	10/03/01	Newcastle	Sunderland	Shrewsbury	Sampled before all Sunderland isolates in subclade.
X7922_1.34	26/07/01	Papworth	Sheffield	Sheffield	NA
X7922_1.38	28/11/01	Papworth	Ashford	Ashford	NA

X7554_6.68	10/01/02	Dublin	Cork	Cork	NA
X7554_6.93	04/02/02	Cambridge	Cork	Chelmsford	Sampled before all Cork isolates in subclade.
X7414_7.72	17/02/02	West Suffolk	Cambridge	Papworth	Sampled before all Cambridge isolates in subclade.
X7414_7.74	14/03/02	West Suffolk	Cambridge	West Suffolk	Sampled before all Cambridge isolates in subclade.
X7414_7.80	17/03/02	Shrewsbury	Chester	Chester	NA
X7414_7.56	20/03/02	Glasgow Royal Infirmary	Kirkcaldy	Dundee	Sampled before all Kirkcaldy isolates in subclade.
X8140_1.67	02/05/02	Papworth	Cambridge	Papworth	Sampled before all Cambridge isolates in subclade.
X7922_1.47	28/05/02	Papworth	Bangor	UCL	Sampled before all Bangor isolates in subclade.
X7414_7.55	07/06/02	Newcastle	Sunderland	Sunderland	NA
X7712_8.38	02/01/03	UCL	Papworth	Papworth	NA
X7414_8.32	15/01/03	Cambridge	Papworth	Papworth	NA
X7712_8.12	07/02/03	Shrewsbury	Bangor	Papworth	Sampled before all Bangor isolates in subclade.
X7922_1.62	23/02/03	Papworth	West Suffolk	West Suffolk	NA
X7712_8.32	04/04/03	Sunderland	Leicester	Coventry	Sampled before all Leicester isolates in subclade.
X7922_1.68	12/06/03	Papworth	Cambridge	Papworth	Sampled before all Cambridge isolates in

					subclade.
X8728_5.23	15/08/03	Kirkcaldy	Altnaegelvin	Altnaegelvin	NA
X7712_8.73	06/01/04	Cambridge	Papworth	Papworth	NA
X7414_8.62	09/01/04	Leicester	Chester	Chester	NA
X7712_8.69	17/01/04	Bristol Royal Infirmary	Manchester	Manchester	NA
X7414_8.80	18/01/04	Shrewsbury	Birmingham	Birmingham	NA
X7414_8.87	20/01/04	Sunderland	Newcastle	Newcastle	NA
X7521_5.6	22/01/04	UCL	Truro	London St. Mary	Sampled before all Truro isolates in subclade.
X7414_8.48	25/01/04	Coventry	Shrewsbury	Shrewsbury	NA
X7712_8.42	25/01/04	Altnaegelvin	Belfast	Altnaegelvin	Could be considered part of larger subclade with majority from Altnaegelvin.
X8728_5.24	03/02/04	Altnaegelvin	Belfast	Altnaegelvin	Could be considered part of larger subclade with majority from Altnaegelvin.
X7414_8.74	10/02/04	Shrewsbury	Cardiff	Cardiff	NA
X7922_1.86	19/02/04	Papworth	Bangor	Papworth	Could be considered smaller subclade with closest isolate from Papworth.
X7414_8.83	20/02/04	London St. Mary	Coventry	Coventry	NA
X7480_7.29	06/01/05	Shrewsbury	Chester	Chester	NA
X7521_5.30	07/01/05	Belfast	Dublin	Dublin	NA

X7521_5.40	13/01/05	Cambridge	Papworth	Papworth	NA
X7480_7.39	14/01/05	Sunderland	Newcastle	Newcastle	NA
X7480_7.13	18/01/05	Newcastle	Sunderland	Newcastle	Could be considered smaller subclade of Newcastle only isolates.
X7521_5.13	24/01/05	Altnaegelvin	Papworth	Papworth	NA
X7480_7.24	11/02/05	Manchester	Southampton	Southampton	NA
X7480_7.10	23/02/05	Newcastle	Sunderland	Newcastle	Could be considered smaller subclade of Newcastle only isolates.
X8140_1.16	20/04/05	Papworth	Cambridge	Papworth	Mistakenly assigned to subclade, should actually be a solo isolate.
X8140_1.25	11/08/05	Papworth	West Suffolk	West Suffolk	NA
X8140_1.26	26/08/05	Papworth	West Suffolk	West Suffolk	NA
X8140_1.34	30/12/05	Papworth	Cambridge	Papworth	Only 2 isolates sampled before in subclade and closest is from Papworth.
X7480_7.64	12/01/06	Dublin	Cork	Dublin	Could be considered separate small subclade with closest isolate from Dublin.
X7480_7.65	15/01/06	Dublin	Kirkcaldy	Kirkcaldy	NA
X7521_6.9	16/01/06	UCL	Truro	London St. Mary	Sampled before all Truro isolates in subclade.
X7521_	21/01/06	Kirkcaldy	Glasgow Royal	Glasgow Royal	NA

5.58			Infirmary	Infirmary	
X7521_5.79	22/01/06	Southampton	Bangor	Papworth	Could be considered smaller subclade with closest isolate from Papworth.
X7480_7.67	23/01/06	Dublin	Cork	Cork	NA
X7480_7.54	06/03/06	Altnaegelvin	Dublin	Dublin	NA
X7065_8.2	20/07/06	Cambridge	Papworth	Papworth	NA
X7065_8.12	19/09/06	Cambridge	West Suffolk	West Suffolk	NA
X7065_8.15	26/09/06	Cambridge	West Suffolk	West Suffolk	NA
X8140_1.51	19/10/06	Papworth	Cambridge	Cambridge	NA
X8140_1.52	12/12/06	Papworth	Cambridge	Papworth	Could be considered smaller subclade with closest isolate from Papworth.
X7521_6.50	01/01/07	Coventry	Chester	Coventry	Could be considered part of larger Coventry subclade.
X7521_6.25	18/01/07	Dublin	Cork	Papworth	Sampled before all Cork isolates in subclade.
X7521_6.27	22/01/07	Belfast	Dublin	Dublin	NA
X7469_7.2	31/01/07	West Suffolk	Cambridge	Cambridge	NA
X7521_6.72	31/01/07	Shrewsbury	Bangor	Bangor	NA
X7521_6.76	18/02/07	Southampton	Truro	Truro	NA
X7065_8.42	19/02/07	Cambridge	Papworth	Papworth	NA
X7065_8.43	26/02/07	Cambridge	Bangor	Papworth	Could be considered

					smaller subclade with closest isolate from Papworth.
X7521_6.69	26/02/07	Manchester	Southampton	Southampton	NA
X7521_6.58	18/03/07	Glasgow Victoria	Glasgow Royal Infirmary	Glasgow Royal Infirmary	NA
X8140_1.55	14/08/07	Papworth	Southampton	Southampton	NA
X7922_1.5	28/08/07	Papworth	Dublin	Belfast	Could be considered separate small subclade with closest isolate from Belfast.
X8140_1.57	14/09/07	Papworth	Cambridge	Cambridge	NA
X7065_8.76	17/12/07	Cambridge	Bristol Royal Infirmary	Bristol Royal Infirmary	NA
X7564_8.12	06/01/08	Manchester	Newcastle	Manchester	Sampled before all Newcastle isolates in subclade.
X7469_7.39	07/01/08	Cardiff	Southampton	Southampton	NA
X7469_7.84	23/01/08	Southampton	Papworth	Papworth	NA
X7469_7.19	05/02/08	Dublin	Cork	Cork	NA
X7469_7.20	05/02/08	Dublin	Cork	Dublin	Could be considered part of smaller subclade with closest isolate from Dublin.
X7469_7.7	16/02/08	Ashford	Cork	Cork	NA
X7469_7.9	22/02/08	Ashford	Truro	Truro	NA
X7480_8.19	26/02/08	UCL	Cork	Cork	NA
X7469_7.37	01/03/08	Cambridge	West Suffolk	West Suffolk	NA

X7469_7.68	10/03/08	Glasgow Victoria	Glasgow South General	Glasgow Victoria	Sampled before all Glasgow South General isolates in subclade.
X7469_7.79	18/03/08	Sheffield	Manchester	Manchester	NA
X7469_7.26	25/03/08	Dublin	Cork	Cork	NA
X7469_7.15	08/04/08	Bangor	Manchester	Manchester	NA
X7065_8.87	08/05/08	Cambridge	Papworth	Papworth	NA
X7065_8.91	20/07/08	Cambridge	Papworth	Papworth	NA
X7469_7.76	02/08/08	Manchester	Newcastle	Newcastle	NA
X7748_6.64	01/01/09	West Suffolk	Sunderland	Sunderland	NA
X7480_8.38	08/01/09	Dublin	Cork	Dublin	Could be considered part of smaller subclade with closest isolate from Dublin.
X7748_6.65	10/01/09	West Suffolk	Papworth	Papworth	NA
X7748_6.35	14/01/09	Manchester	Ashford	Ashford	NA
X7480_8.23	26/01/09	Antrim	Belfast	Belfast	NA
X7480_8.24	28/01/09	Antrim	Glasgow Royal Infirmary	Glasgow Royal Infirmary	NA
X7083_1.8	31/01/09	Cambridge	Papworth	Papworth	NA
X7480_8.61	03/02/09	Cardiff	Bristol Royal Infirmary	Bristol Royal Infirmary	NA
X7480_8.27	19/03/09	Antrim	Belfast	Belfast	NA
X7083_1.13	24/03/09	Cambridge	Papworth	Papworth	NA
X7480_8.68	27/03/09	Cork	Dublin	Dublin	NA
X7480_8.41	31/03/09	Dublin	Cork	Dublin	Could be considered part

					of smaller subclade with closest isolate from Dublin.
X7083_1.16	04/05/09	Cambridge	Papworth	Papworth	NA
X7480_8.82	11/05/09	Inverness	Wishaw	Sunderland	Sampled before all Wishaw isolates in subclade.
X7480_8.64	26/05/09	Chester	Sheffield	Sheffield	NA
X7083_1.22	02/08/09	Cambridge	Papworth	Papworth	NA
X7083_1.24	28/10/09	Cambridge	Truro	Truro	NA
X7564_8.93	06/01/10	Bristol Southmead	Bristol Royal Infirmary	Bristol Royal Infirmary	NA
X7564_8.21	12/01/10	Belfast	Glasgow Royal Infirmary	Glasgow Royal Infirmary	NA
X7564_8.22	21/01/10	Belfast	Dublin	Belfast	Could be considered separate small subclade with closest isolate from Belfast.
X7564_8.31	23/01/10	Cardiff	Truro	Cambridge	Could be considered separate small subclade with closest isolate from Cambridge.
X7564_8.80	26/01/10	Galway	Cork	Cork	NA
X7564_8.85	26/01/10	Wishaw	Glasgow Royal Infirmary	Glasgow Royal Infirmary	NA
X7083_1.28	27/01/10	Cambridge	Papworth	Papworth	NA
X7083_1.29	07/02/10	Cambridge	Papworth	Cambridge	Could be considered a separate small subclade with closest isolate from Cambridge.

X7564_8.54	09/02/10	Leicester	Southampton	Southampton	NA
X7564_8.65	17/02/10	Shrewsbury	Bangor	Bangor	NA
X7564_8.37	20/02/10	Chelsea	London St. Mary	London St. Mary	NA
X7564_8.88	06/04/10	Wishaw	Glasgow Royal Infirmary	Glasgow Royal Infirmary	NA
X7083_1.30	12/05/10	Cambridge	Papworth	Papworth	NA
X7083_1.35	26/06/10	Cambridge	Papworth	Papworth	NA
X7748_6.80	29/06/10	Bangor	West Suffolk	West Suffolk	NA
X7915_6.10	10/07/10	Colchester	Cambridge	Cambridge	NA
X7083_1.38	18/08/10	Cambridge	Papworth	Papworth	NA
X7564_8.47	17/09/10	Edinburgh Royal Infirmary	Inverness	Inverness	NA
X7564_8.48	16/12/10	Edinburgh Royal Infirmary	Inverness	Edinburgh Royal Infirmary	Could be considered a separate small subclade with closest isolate from Edinburgh Royal Infirmary.
X7564_8.49	16/12/10	Edinburgh Royal Infirmary	Kirkcaldy	Kirkcaldy	NA

Supplementary Table C2. The summary of all 90 isolates sampled in 2010 with the predicted origin hospital from three independent investigators using the TAPO compared to the posited origin from the SnAPO method. Each investigator was also asked to indicate how confident they were with their posited hospital origins. Furthermore, I have included the SnAPO Diagnostic Origin Value (DOV) for the posited origin location.

Strain Code	Sampling Date	Sampling Location	SnAPO	SnAPO DOV	TAPO 1	TAPO 1 Confidence	TAPO 2	TAPO 2 Confidence	TAPO 3	TAPO 3 Confidence
X7564_8.19	01/01/2010	Belfast	Antrim	0.29	Belfast	High	Antrim	Low	Antrim	Low
X7564_8.91	01/01/2010	Wrexham	Chester	0.536	NA	NA	NA	NA	NA	NA
X7564_8.25	03/01/2010	Cardiff	Cardiff	0.803	Cardiff	High	Cardiff	High	Cardiff	High
X7564_8.92	03/01/2010	Bristol Southmead	Bristol Royal Infirmary	0.512	Bristol Royal Infirmary	Low	Bristol Royal Infirmary	Low	Bristol Royal Infirmary	Low
X7564_8.69	04/01/2010	Southampton	Southampton	0.661	Southampton	High	Southampton	High	Southampton	Low
X7564_8.20	05/01/2010	Belfast	Belfast	0.241	Belfast	High	Belfast	Low	Belfast	High
X7564_8.26	05/01/2010	Cardiff	Leicester	0.596	Leicester	Low	NA	NA	NA	NA
X7564_8.44	06/01/2010	Coventry	Coventry	0.499	NA	NA	Coventry	Low	NA	NA
X7564_8.93	06/01/2010	Bristol Southmead	Bristol Royal Infirmary	0.338	Bristol Royal Infirmary	Low	Bristol Royal Infirmary	High	Bristol Royal Infirmary	Low
X7564_8.50	07/01/2010	Newcastle	Newcastle	0.644	Newcastle	Low	NA	NA	Newcastle	Low
X7564_8.66	07/01/2010	Glasgow Southern General	Glasgow Southern General	0.406	Glasgow Southern General	Low	Glasgow Southern General	Low	NA	NA
X7564_8.35	09/01/2010	Chelsea	London St. Mary	0.307	NA	NA	NA	NA	London St. Mary	Low
X7564_8.21	12/01/2010	Belfast	Glasgow Royal Infirmary	0.406	Antrim	Low	Antrim	Low	Glasgow Royal Infirmary	Low
X7564_8.27	12/01/2010	Cardiff	Cardiff	0.416	Cardiff	High	Cardiff	High	Cardiff	High
X7564_8.61	13/01/2010	London Northwick	Cambridge	0.125	London Northwick	Low	London Northwick	High	London Northwick	High
X7748_6.69	13/01/2010	Antrim	Altnaegelvin	0.172	Altnaegelvin	Low	Altnaegelvin	Low	Altnaegelvin	Low
X7748_6.70	13/01/2010	Antrim	Antrim	0.381	Altnaegelvin	Low	Altnaegelvin	Low	Altnaegelvin	Low
X7564_8.28	14/01/2010	Cardiff	Cardiff	0.437	Cardiff	High	Shrewsbury	Low	Cardiff	High
X7564_8.52	14/01/2010	Leicester	Leicester	0.522	Leicester	Low	Leicester	Low	Leicester	Low
X7564_8.29	15/01/2010	Cardiff	Cardiff	0.611	Cardiff	High	Cardiff	High	Cardiff	High
X7564_8.30	17/01/2010	Cardiff	Cardiff	0.594	Cardiff	High	Cardiff	High	Cardiff	High
X7564_8.74	18/01/2010	Sunderland	Sunderland	0.585	Sunderland	High	Sunderland	High	Sunderland	Low
X7564_8.84	20/01/2010	Wishaw	Inverness	0.532	Inverness	Low	Inverness	Low	Newcastle	Low
X7564_8.22	21/01/2010	Belfast	Belfast	0.361	Belfast	Low	NA	NA	Dublin	Low
X7564_8.31	23/01/2010	Cardiff	Cambridge	0.399	Truro	High	Cambridge	Low	Truro	High
X7564_8.83	24/01/2010	West Suffolk	Papworth	0.51	NA	NA	Papworth	Low	Papworth	Low
X7564_8.15	26/01/2010	Dublin	Dublin	0.509	Dublin	High	Dublin	High	Dublin	High
X7564_8.80	26/01/2010	Galway	Cork	0.34	Cork	Low	Dublin	Low	Dublin	High
X7564_8.85	26/01/2010	Wishaw	Glasgow Royal Infirmary	0.207	Glasgow	High	Glasgow Royal Infirmary	High	Glasgow Royal Infirmary	Low
X7083_1.28	27/01/2010	Cambridge	Papworth	0.485	NA	NA	Papworth	Low	NA	NA
X7564_8.53	28/01/2010	Leicester	Leicester	0.504	Leicester	High	Leicester	High	Leicester	High
X7748_6.77	29/01/2010	Bangor	Bangor	0.547	Bangor	High	Bangor	High	Bangor	High
X7564_8.75	05/02/2010	Sunderland	Newcastle	0.267	Newcastle	Low	Newcastle	Low	Newcastle	Low
X7083_1.29	07/02/2010	Cambridge	Cambridge	0.575	NA	NA	Papworth	Low	NA	NA
X7564_8.79	07/02/2010	Ulster	Belfast	0.292	Belfast	Low	Antrim	Low	Belfast	Low
X7564_8.76	08/02/2010	Sunderland	Sunderland	0.597	Sunderland	High	Sunderland	High	Sunderland	High
X7564_8.54	09/02/2010	Leicester	Southampton	0.626	Southampton	High	Southampton	High	Southampton	Low
X7564_8.68	09/02/2010	Glasgow Southern General	UCL	0.35	UCL	High	UCL	High	UCL	High
X7564_8.70	09/02/2010	Southampton	Southampton	0.776	Southampton	High	Southampton	High	Southampton	Low
X7748_6.66	09/02/2010	York	Cambridge	0.263	NA	NA	Southampton	Low	Southampton	Low
X7748_6.67	11/02/2010	York	Sheffield	0.236	NA	NA	NA	NA	NA	NA
X7564_8.86	13/02/2010	Wishaw	Manchester	0.273	NA	NA	Birmingham	Low	NA	NA
X7748_6.72	16/02/2010	UCL	London St. Mary	0.223	Dublin	Low	NA	NA	Dublin	Low
X7564_8.65	17/02/2010	Shrewsbury	Bangor	0.475	Bangor	High	Bangor	High	Bangor	High
X7564_8.82	18/02/2010	Galway	Cork	0.539	Cork	Low	NA	NA	Cork	Low

X7564_8.37	20/02/2010	Chelsea	London St. Mary	0.575	London St. Mary	High	London St. Mary	High	London St. Mary	High
X7748_6.73	20/02/2010	UCL	London St. Mary	0.302	London St. Mary	Low	London St. Mary	Low	NA	NA
X7564_8.17	21/02/2010	Dublin	Dublin	0.265	Dublin	Low	Dublin	Low	Dublin	Low
X7564_8.18	26/02/2010	Dublin	Dublin	0.549	Dublin	High	Dublin	High	Dublin	High
X7748_6.75	01/03/2010	UCL	London St. Mary	0.244	Colchester	Low	NA	NA	London St. Mary	Low
X7564_8.77	02/03/2010	Truro	Truro	0.54	Truro	High	Truro	High	Truro	High
X7564_8.71	06/03/2010	London St. Barts	UCL	0.263	UCL	High	UCL	High	UCL	High
X7564_8.24	20/03/2010	Cambridge	Papworth	0.241	Papworth	Low	Papworth	Low	Papworth	Low
X7564_8.64	27/03/2010	Inverness	Inverness	0.826	Inverness	High	Inverness	High	Inverness	High
X7564_8.16	04/04/2010	Dublin	Dublin	0.507	Dublin	High	Dublin	High	Dublin	High
X7564_8.38	06/04/2010	Cork	Dublin	0.55	Dublin	High	Dublin	High	Dublin	High
X7564_8.88	06/04/2010	Wishaw	Glasgow Royal Infirmary	0.207	Glasgow	High	Glasgow Royal Infirmary	High	Glasgow Royal Infirmary	Low
X7564_8.43	08/04/2010	Cork	Cork	0.495	Cork	High	Cork	High	Cork	High
X7564_8.42	11/04/2010	Cork	Dublin	0.4	Dublin	High	Dublin	High	Dublin	High
X7564_8.40	13/04/2010	Cork	Cork	0.552	Cork	High	Cork	High	Cork	High
X7915_6.9	14/04/2010	Colchester	London St. Mary	0.74	London St. Mary	Low	NA	NA	NA	NA
X7564_8.89	16/04/2010	Wishaw	Wishaw	0.511	NA	NA	Birmingham	Low	NA	NA
X7564_8.32	17/04/2010	Chester	Chester	0.575	NA	NA	Chester	Low	Chester	Low
X7564_8.90	21/04/2010	Wrexham	Shrewsbury	0.237	NA	NA	NA	NA	Shrewsbury	Low
X7564_8.33	29/04/2010	Chester	Chester	0.538	NA	NA	NA	NA	NA	NA
X7083_1.30	12/05/2010	Cambridge	Papworth	0.42	Papworth	High	Papworth	Low	Papworth	Low
X7564_8.34	18/05/2010	Chester	Chester	0.273	Chester	High	Chester	High	NA	NA
X7083_1.31	20/05/2010	Cambridge	Papworth	0.487	Papworth	High	Papworth	High	Papworth	High
X8728_5.51	24/05/2010	Colchester	UCL	0.267	Dublin	Low	NA	NA	Dublin	Low
X7083_1.32	07/06/2010	Cambridge	Cambridge	0.473	Cambridge	High	Cambridge	High	Cambridge	High
X7083_1.34	25/06/2010	Cambridge	UCL	0.26	UCL	High	UCL	High	UCL	High
X7083_1.35	26/06/2010	Cambridge	Papworth	0.567	Cambridge	Low	Cambridge	Low	Cambridge	High
X7748_6.80	29/06/2010	Bangor	West Suffolk	0.315	Papworth	Low	Papworth	Low	Papworth	Low
X7083_1.36	30/06/2010	Cambridge	Cambridge	0.741	Cambridge	High	Cambridge	High	Cambridge	High
X7564_8.56	30/06/2010	North Devon	Chester	0.542	NA	NA	Chester	Low	Chester	Low
X7915_6.10	10/07/2010	Colchester	Cambridge	0.442	Papworth	Low	Papworth	Low	Papworth	Low
X7083_1.38	18/08/2010	Cambridge	Papworth	0.633	Papworth	High	Papworth	High	Papworth	High
X7564_8.58	28/08/2010	North Devon	Cardiff	0.415	Cardiff	High	Cardiff	High	Cardiff	High
X8728_5.39	12/09/2010	Edinburgh Royal Infirmary	Kirkcaldy	0.135	NA	NA	Glasgow Victoria	Low	NA	NA
X7083_1.39	14/09/2010	Cambridge	Cambridge	0.46	Cambridge	Low	Cambridge	Low	Cambridge	High
X7564_8.47	17/09/2010	Edinburgh Royal Infirmary	Inverness	0.7	Inverness	High	Inverness	High	Inverness	High
X8728_5.38	19/09/2010	Edinburgh Royal Infirmary	Kirkcaldy	0.177	NA	NA	Kirkcaldy	Low	Glasgow Royal Infirmary	Low
X7564_8.73	02/10/2010	London St. Barts	UCL	0.251	UCL	High	UCL	High	UCL	High
X7915_6.13	08/10/2010	Colchester	Norfolk	0.501	NA	NA	Norfolk	Low	Norfolk	Low
X7564_8.55	10/10/2010	Manchester	Manchester	0.576	Manchester	High	Manchester	Low	NA	NA
X7564_8.81	10/10/2010	Galway	Altnaegelvin	0.587	Altnaegelvin	Low	NA	NA	Altnaegelvin	Low
X7564_8.87	10/10/2010	Wishaw	Wishaw	0.781	Inverness	Low	Inverness	Low	Newcastle	Low
X7915_6.14	30/10/2010	Colchester	Papworth	0.158	NA	NA	NA	NA	Belfast	Low
X7564_8.48	16/12/2010	Edinburgh Royal Infirmary	Edinburgh Royal Infirmary	0.477	Inverness	High	Inverness	High	Inverness	High
X7564_8.49	16/12/2010	Edinburgh Royal Infirmary	Kirkcaldy	0.497	Kirkcaldy	Low	Kirkcaldy	High	Kirkcaldy	High

Supplementary Table C3. The summary of all 90 isolates sampled in 2010 with the predicted origin Referral Cluster (RC) from three independent investigators using the TAPO compared to the posited origin from the SnAPO method. Each investigator was also asked to indicate how confident they were with their posited RC origins. Furthermore, I have included the SnAPO Diagnostic Origin Value (DOV) for the posited origin location.

Strain Code	Sampling Date	Sampling Location	SnAPO	SnAPO DOV	TAPO 1	TAPO 1 Confidence	TAPO 2	TAPO 2 Confidence	TAPO 3	TAPO 3 Confidence
X7564_8.19	01/01/2010	RC14	RC14	0.568	RC14	High	RC14	High	RC14	High
X7564_8.91	01/01/2010	RC16	RC7	0.536	NA	NA	NA	NA	NA	NA
X7564_8.25	03/01/2010	RC16	RC16	0.816	RC16	High	RC16	High	RC16	High
X7564_8.92	03/01/2010	RC2	RC2	0.679	RC2	High	RC2	High	RC2	High
X7564_8.69	04/01/2010	RC5	RC5	0.661	RC5	High	RC5	High	RC5	Low
X7564_8.20	05/01/2010	RC14	RC14	0.334	RC14	High	RC14	High	RC14	High
X7564_8.26	05/01/2010	RC16	RC9	0.596	RC9	Low	NA	NA	NA	NA
X7564_8.44	06/01/2010	RC6	RC6	0.506	NA	NA	RC6	Low	NA	NA
X7564_8.93	06/01/2010	RC2	RC2	0.544	RC2	High	RC2	High	RC2	Low
X7564_8.50	07/01/2010	RC10	RC10	0.672	RC10	Low	NA	NA	RC10	Low
X7564_8.66	07/01/2010	RC15	RC15	0.774	RC15	High	RC15	High	RC15	High
X7564_8.35	09/01/2010	RC1	RC1	0.348	NA	NA	NA	NA	RC1	Low
X7564_8.21	12/01/2010	RC14	RC15	0.531	RC14	Low	RC14	Low	RC15	Low
X7564_8.27	12/01/2010	RC16	RC16	0.44	RC16	High	RC16	High	RC16	High
X7564_8.61	13/01/2010	RC1	RC1	0.317	RC1	High	RC1	High	RC1	High
X7748_6.69	13/01/2010	RC14	RC14	0.242	RC14	High	RC14	High	RC14	High
X7748_6.70	13/01/2010	RC14	RC14	0.508	RC14	High	RC15	High	RC14	High
X7564_8.28	14/01/2010	RC16	RC16	0.448	RC16	High	RC6	Low	RC16	High
X7564_8.52	14/01/2010	RC9	RC9	0.522	RC9	Low	RC9	Low	RC9	Low
X7564_8.29	15/01/2010	RC16	RC16	0.622	RC16	High	RC16	High	RC16	High
X7564_8.30	17/01/2010	RC16	RC16	0.608	RC16	High	RC16	High	RC16	High
X7564_8.74	18/01/2010	RC10	RC10	0.593	RC10	High	RC10	High	RC10	Low
X7564_8.84	20/01/2010	RC15	RC15	0.547	RC15	Low	RC15	Low	RC10	Low
X7564_8.22	21/01/2010	RC14	RC14	0.485	RC14	Low	NA	NA	RC13	Low
X7564_8.31	23/01/2010	RC16	RC8	0.424	RC2	High	RC8	Low	RC2	High
X7564_8.83	24/01/2010	RC8	RC8	0.556	RC8	Low	RC8	High	RC8	High
X7564_8.15	26/01/2010	RC13	RC13	0.577	RC13	High	RC13	High	RC13	High
X7564_8.80	26/01/2010	RC13	RC13	0.494	RC13	High	RC13	High	RC13	High
X7564_8.85	26/01/2010	RC15	RC15	0.264	RC15	High	RC15	High	RC15	Low
X7083_1.28	27/01/2010	RC8	RC8	0.587	RC8	Low	RC8	Low	NA	NA
X7564_8.53	28/01/2010	RC9	RC9	0.505	RC9	High	RC9	High	RC9	High
X7748_6.77	29/01/2010	RC16	RC16	0.552	RC16	High	RC16	High	RC16	High
X7564_8.75	05/02/2010	RC10	RC10	0.454	RC10	High	RC10	High	RC10	High
X7083_1.29	07/02/2010	RC8	RC8	0.811	RC8	Low	RC8	Low	NA	NA
X7564_8.79	07/02/2010	RC14	RC14	0.63	RC14	High	RC14	High	RC14	High
X7564_8.76	08/02/2010	RC10	RC10	0.704	RC10	High	RC10	High	RC10	High
X7564_8.54	09/02/2010	RC9	RC5	0.626	RC5	High	RC5	High	RC5	Low
X7564_8.68	09/02/2010	RC15	RC4	0.361	RC4	High	RC4	High	RC4	High
X7564_8.70	09/02/2010	RC5	RC5	0.776	RC5	High	RC5	High	RC5	Low
X7748_6.66	09/02/2010	RC3	RC8	0.308	NA	NA	RC5	Low	RC5	Low
X7748_6.67	11/02/2010	RC3	RC8	0.259	NA	NA	NA	NA	NA	NA
X7564_8.86	13/02/2010	RC15	RC12	0.273	NA	NA	RC6	Low	NA	NA
X7748_6.72	16/02/2010	RC4	RC1	0.277	RC13	Low	NA	NA	RC13	Low
X7564_8.65	17/02/2010	RC6	RC16	0.499	RC16	High	RC16	High	RC16	High
X7564_8.82	18/02/2010	RC13	RC13	0.594	RC13	Low	NA	NA	RC13	Low

X7564_8.37	20/02/2010	RC1	RC1	0.603	RC1	High	RC1	High	RC1	High
X7748_6.73	20/02/2010	RC4	RC1	0.362	RC1	Low	RC1	Low	RC1	High
X7564_8.17	21/02/2010	RC13	RC13	0.384	RC13	Low	RC13	Low	RC13	Low
X7564_8.18	26/02/2010	RC13	RC13	0.601	RC13	High	RC13	High	RC13	High
X7748_6.75	01/03/2010	RC4	RC1	0.33	RC4	Low	NA	NA	RC1	Low
X7564_8.77	02/03/2010	RC2	RC2	0.54	RC2	High	RC2	High	RC2	High
X7564_8.71	06/03/2010	RC4	RC4	0.271	RC4	High	RC4	High	RC4	High
X7564_8.24	20/03/2010	RC8	RC8	0.505	RC8	High	RC8	High	RC8	High
X7564_8.64	27/03/2010	RC15	RC15	0.828	RC15	High	RC15	High	RC15	High
X7564_8.16	04/04/2010	RC13	RC13	0.566	RC13	High	RC13	High	RC13	High
X7564_8.38	06/04/2010	RC13	RC13	0.787	RC13	High	RC13	High	RC13	High
X7564_8.88	06/04/2010	RC15	RC15	0.311	RC15	High	RC15	High	RC15	Low
X7564_8.43	08/04/2010	RC13	RC13	0.679	RC13	High	RC13	High	RC13	High
X7564_8.42	11/04/2010	RC13	RC13	0.769	RC13	High	RC13	High	RC13	High
X7564_8.40	13/04/2010	RC13	RC13	0.688	RC13	High	RC13	High	RC13	High
X7915_6.9	14/04/2010	RC4	RC1	0.749	RC1	Low	NA	NA	NA	NA
X7564_8.89	16/04/2010	RC15	RC15	0.513	NA	NA	RC6	Low	NA	NA
X7564_8.32	17/04/2010	RC7	RC7	0.575	RC2	Low	RC7	Low	RC7	Low
X7564_8.90	21/04/2010	RC16	RC6	0.245	NA	NA	NA	NA	RC6	Low
X7564_8.33	29/04/2010	RC7	RC7	0.538	NA	NA	NA	NA	NA	NA
X7083_1.30	12/05/2010	RC8	RC8	0.509	RC8	High	RC8	Low	RC8	Low
X7564_8.34	18/05/2010	RC7	RC7	0.273	RC7	High	RC7	High	NA	NA
X7083_1.31	20/05/2010	RC8	RC8	0.737	RC8	High	RC8	High	RC8	High
X8728_5.51	24/05/2010	RC4	RC4	0.293	RC13	Low	NA	NA	RC13	Low
X7083_1.32	07/06/2010	RC8	RC8	0.842	RC8	High	RC8	High	RC8	High
X7083_1.34	25/06/2010	RC8	RC4	0.346	RC4	High	RC4	High	RC4	High
X7083_1.35	26/06/2010	RC8	RC8	0.921	RC8	High	RC8	High	RC8	High
X7748_6.80	29/06/2010	RC16	RC8	0.526	RC8	High	RC8	High	RC8	Low
X7083_1.36	30/06/2010	RC8	RC8	0.937	RC8	High	RC8	High	RC8	High
X7564_8.56	30/06/2010	RC2	RC7	0.542	RC2	Low	RC7	Low	RC7	Low
X7915_6.10	10/07/2010	RC4	RC8	0.642	RC8	High	RC8	High	RC8	High
X7083_1.38	18/08/2010	RC8	RC8	0.912	RC8	High	RC8	High	RC8	High
X7564_8.58	28/08/2010	RC2	RC16	0.498	RC16	High	RC16	High	RC16	High
X8728_5.39	12/09/2010	RC15	RC15	0.387	RC15	High	RC15	High	RC15	High
X7083_1.39	14/09/2010	RC8	RC8	0.758	RC8	High	RC8	High	RC8	High
X7564_8.47	17/09/2010	RC15	RC15	0.703	RC15	High	RC15	High	RC15	High
X8728_5.38	19/09/2010	RC15	RC15	0.348	RC15	High	RC15	High	RC15	Low
X7564_8.73	02/10/2010	RC4	RC4	0.333	RC4	High	RC4	High	RC4	High
X7915_6.13	08/10/2010	RC4	RC8	0.68	NA	NA	RC8	Low	RC8	Low
X7564_8.55	10/10/2010	RC12	RC12	0.576	RC12	High	RC12	High	NA	NA
X7564_8.81	10/10/2010	RC13	RC14	0.665	RC14	Low	NA	NA	RC14	Low
X7564_8.87	10/10/2010	RC15	RC15	0.876	RC15	Low	RC16	Low	RC10	Low
X7915_6.14	30/10/2010	RC4	RC8	0.298	NA	NA	NA	NA	RC14	Low
X7564_8.48	16/12/2010	RC15	RC15	0.836	RC15	High	RC15	High	RC15	High
X7564_8.49	16/12/2010	RC15	RC15	0.522	RC15	High	RC15	High	RC15	High

Appendix D

Supplementary Table D1. The origin locations posited by SnAPO and Bayesian inference, with the values of those posited origins. The SnAPO values show greater variation than the Bayesian values. If the origin posited by both methods concurs then the values posited by both should be considered, since agreement does not automatically mean that it is the true origin.

Strain Code	Sampled	SnAPO Origin	Bayesian Origin	SnAPO Maximum Value	Bayes Maximum Value
X7564_8.19	Belfast	Antrim	Belfast	29.01	100.00
X7564_8.91	Wrexham	Chester	Chester	53.62	100.00
X7564_8.25	Cardiff	Cardiff	Cardiff	80.30	100.00
X7564_8.92	Bristol Southmead	Bristol Royal Infirmary	Bristol Royal Infirmary	51.23	99.92
X7564_8.69	Southampton	Southampton	Southampton	66.10	100.00
X7564_8.20	Belfast	Belfast	Belfast	24.10	100.00
X7564_8.26	Cardiff	Leicester	Leicester	59.59	100.00
X7564_8.44	Coventry	Coventry	Coventry	49.95	100.00
X7564_8.93	Bristol Southmead	Bristol Royal Infirmary	Bristol Royal Infirmary	33.81	99.99
X7564_8.50	Newcastle	Newcastle	Newcastle	64.40	100.00
X7564_8.66	Glasgow South General	Glasgow South General	Glasgow South General	40.58	100.00
X7564_8.35	Chelsea	London St. Mary	London St. Mary	30.70	100.00
X7564_8.78	Newcastle	Belfast	Belfast	24.42	99.92
X7564_8.21	Belfast	Glasgow Royal Infirmary	Glasgow Royal Infirmary	40.63	100.00
X7564_8.27	Cardiff	Cardiff	Cardiff	41.59	100.00
X7564_8.61	London Northwick	Cambridge	London St. Mary	12.50	98.63
X7748_6.69	Antrim	Altnaegelvin	Sunderland	17.20	99.56
X7748_6.70	Antrim	Antrim	Sunderland	38.09	99.56
X7564_8.28	Cardiff	Cardiff	Cardiff	43.70	100.00
X7564_8.52	Leicester	Leicester	Leicester	52.16	100.00
X7564_8.29	Cardiff	Cardiff	Cardiff	61.05	100.00

X7564_8.30	Cardiff	Cardiff	Cardiff	59.39	100.00
X7564_8.74	Sunderland	Sunderland	Sunderland	58.54	100.00
X7564_8.84	Wishaw	Inverness	Inverness	53.20	68.82
X7564_8.22	Belfast	Belfast	Dublin	36.12	95.52
X7564_8.31	Cardiff	Cambridge	Cardiff	39.93	98.33
X7564_8.83	West Suffolk	Papworth	Papworth	50.96	99.99
X7564_8.15	Dublin	Dublin	Dublin	50.89	100.00
X7564_8.80	Galway	Cork	Cork	34.04	100.00
X7564_8.85	Wishaw	Glasgow Royal Infirmary	Glasgow Royal Infirmary	20.73	99.92
X7083_1.28	Cambridge	Papworth	Papworth	48.55	100.00
X7564_8.53	Leicester	Leicester	Leicester	50.44	100.00
X7748_6.77	Bangor	Bangor	Bangor	54.66	100.00
X7564_8.75	Sunderland	Newcastle	Sunderland	26.68	99.74
X7083_1.29	Cambridge	Cambridge	Papworth	57.49	99.71
X7564_8.79	Ulster	Belfast	Belfast	29.21	100.00
X7564_8.76	Sunderland	Sunderland	Sunderland	59.66	100.00
X7564_8.54	Leicester	Southampton	Southampton	62.58	100.00
X7564_8.68	Glasgow South General	UCL	London St. Mary	34.98	96.71
X7564_8.70	Southampton	Southampton	Southampton	77.59	100.00
X7748_6.66	York	Cambridge	Sheffield	26.35	98.63
X7748_6.67	York	Sheffield	Sheffield	23.61	99.92
X7564_8.86	Wishaw	Manchester	Manchester	27.28	99.82
X7748_6.72	UCL	London St. Mary	London St. Mary	22.28	98.94
X7564_8.65	Shrewsbury	Bangor	Bangor	47.46	100.00
X7564_8.82	Galway	Cork	Cork	53.93	100.00
X7564_8.37	Chelsea	London St. Mary	London St. Mary	57.54	100.00
X7748_6.73	UCL	London St. Mary	London St. Mary	30.23	100.00
X7564_8.17	Dublin	Dublin	Dublin	26.45	99.98
X7564_8.18	Dublin	Dublin	Dublin	54.92	100.00
X7748_6.75	UCL	London St. Mary	London St. Mary	24.36	100.00
X7564_8.77	Truro	Truro	Truro	54.01	72.62
X7564_8.71	London St. Bart's	UCL	UCL	26.25	95.57
X7564_8.24	Cambridge	Papworth	Papworth	24.08	66.26
X7564_8.64	Inverness	Inverness	Inverness	82.61	100.00
X7564_8.16	Dublin	Dublin	Dublin	50.70	100.00
X7564_8.38	Cork	Dublin	Dublin	54.99	100.00

X7564_8.88	Wishaw	Glasgow Royal Infirmary	Glasgow Royal Infirmary	20.69	99.41
X7564_8.43	Cork	Cork	Cork	49.52	100.00
X7564_8.42	Cork	Dublin	Cork	39.97	99.96
X7564_8.40	Cork	Cork	Cork	55.22	100.00
X7915_6.9	Colchester	London St. Mary	London St. Mary	74.03	100.00
X7564_8.89	Wishaw	Wishaw	Manchester	51.09	97.65
X7564_8.32	Chester	Chester	Chester	57.51	100.00
X7564_8.90	Wrexham	Shrewsbury	Shrewsbury	23.67	58.78
X7564_8.33	Chester	Chester	Chester	53.78	99.43
X7083_1.30	Cambridge	Papworth	London St. Mary	42.00	99.72
X7564_8.34	Chester	Chester	Chester	27.33	100.00
X7083_1.31	Cambridge	Papworth	Papworth	48.73	100.00
X8728_5.51	Colchester	UCL	UCL	26.68	83.68
X7083_1.32	Cambridge	Cambridge	Cambridge	47.29	100.00
X7083_1.34	Cambridge	UCL	UCL	25.96	86.25
X7083_1.35	Cambridge	Papworth	Papworth	56.67	100.00
X7748_6.80	Bangor	West Suffolk	West Suffolk	31.47	100.00
X7083_1.36	Cambridge	Cambridge	Cambridge	74.13	100.00
X7564_8.56	North Devon	Chester	Chester	54.16	100.00
X7915_6.10	Colchester	Cambridge	Cambridge	44.19	100.00
X7083_1.38	Cambridge	Papworth	Papworth	63.31	100.00
X7564_8.58	North Devon	Cardiff	Cardiff	41.53	100.00
X8728_5.39	Edinburgh Royal Infirmary	Kirkcaldy	Belfast	13.45	99.96
X7083_1.39	Cambridge	Cambridge	Cambridge	45.96	98.49
X7564_8.47	Edinburgh Royal Infirmary	Inverness	Inverness	69.96	100.00
X8728_5.38	Edinburgh Royal Infirmary	Kirkcaldy	Belfast	17.72	99.97
X7564_8.73	London St. Bart's	UCL	UCL	25.06	94.76
X7915_6.13	Colchester	Norfolk	Cardiff	50.08	79.02
X7564_8.55	Manchester	Manchester	Manchester	57.59	100.00
X7564_8.81	Galway	Altnaegelvin	Altnaegelvin	58.66	100.00
X7564_8.87	Wishaw	Wishaw	Wishaw	78.07	100.00
X7915_6.14	Colchester	Papworth	Cardiff	15.84	51.18
X7564_8.48	Edinburgh Royal Infirmary	Edinburgh Royal Infirmary	Inverness	47.67	100.00

X7564_8.49	Edinburgh Royal Infirmary	Kirkcaldy	Kirkcaldy	49.74	100.00

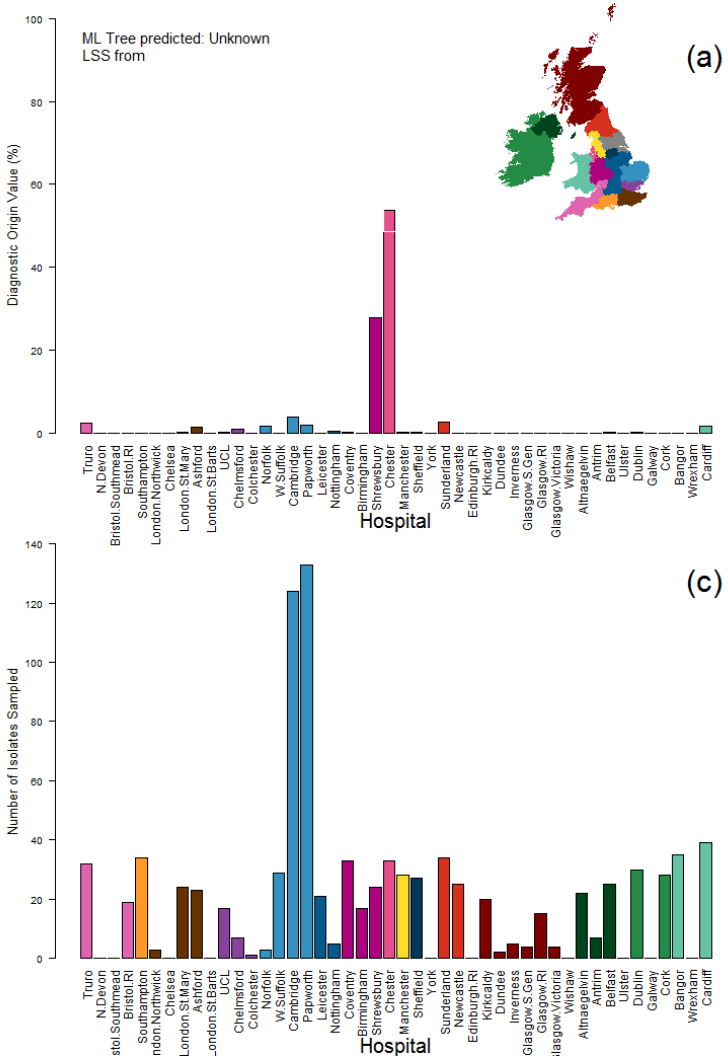
Appendix

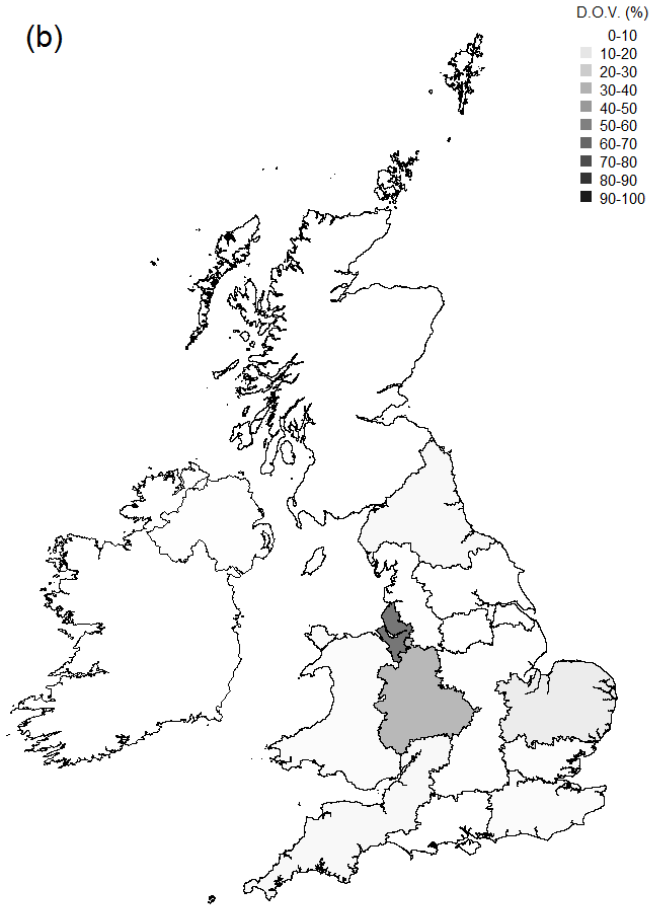
E

Here I provide a select number of examples from the 90 test isolates which were processed using SnAPO, as described in Chapter 4. I chose these examples to exhibit the variety of output possible using SnAPO. Each of the examples have the following layout. The first panel (a) shows the SnAPO output as a DOV percentage. I also display any LSSs the isolate expresses, and the origin location as predicted by TAPO. (b) displays a coarser scale of geographic resolution using RCs. Finally, the sampling effort for each hospital is shown in (c). The bars in (a) and (c) are coloured by the RCs (provided as a graphical legend in a) and are ordered by their geographic proximity. The shading in (b) is split into 10% bins, with the in the top right corner. A brief description of the following examples are provided here:

- Isolate 933 shows two possible origin hospitals in neighbouring RCs.
- Isolate 936 shows one very high peak for a single hospital and RC. This type of output is similar to the majority of the test isolates.
- Isolate 943 shows an ambiguous output, with no one hospital or RC as the obvious origin location.
- Isolate 947 shows a very unclear output, with no location as an obvious origin at either the hospital or RC geographic resolution.
- Isolate 970 shows two possible origin hospitals in neighbouring RCS, though this output is slightly ambiguous due to lower DOVs.
- Isolate 972 shows three possible origin hospitals in three disparate RCs. This output is therefore ambiguous as to the true origin of this isolate.
- Isolate 983 shows two possible origin locations in neighbouring RCs of different countries; England and Wales.
- Isolate 985 shows an ambiguous hospital origin, but a clearer RC origin.
- Isolate 991 shows two very similar origin hospitals, but a very clear RC origin.
- Isolate 1020 shows a very unclear origin location, at both geographic resolutions.

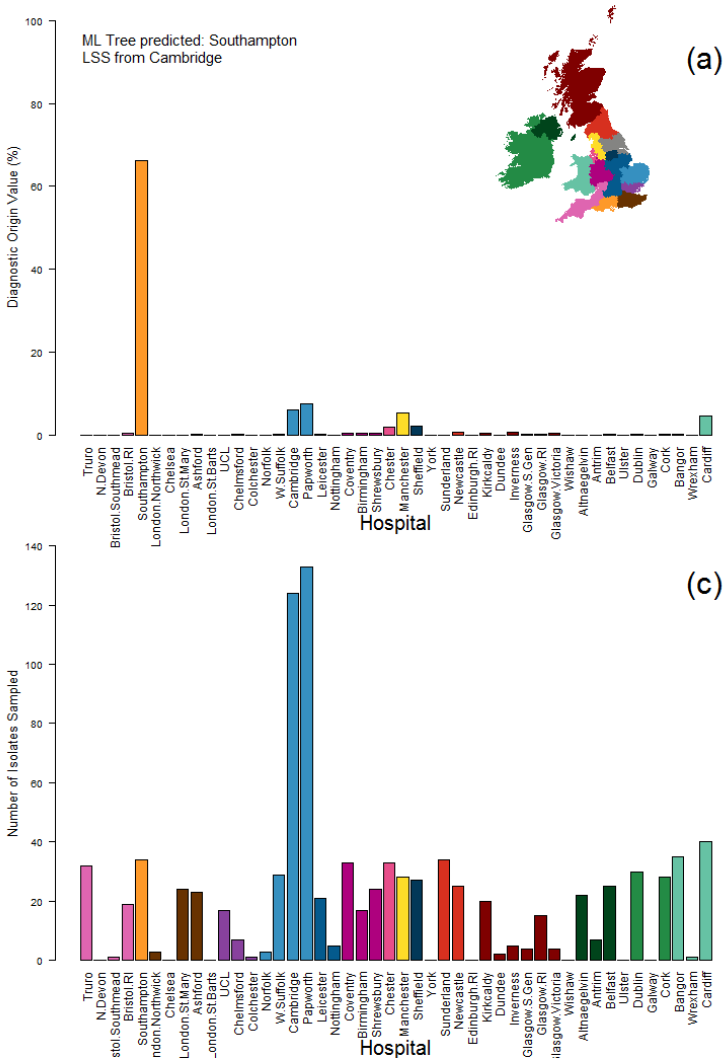
Isolate #933 (X7564_8.91) sampled from Wrexham in 2010

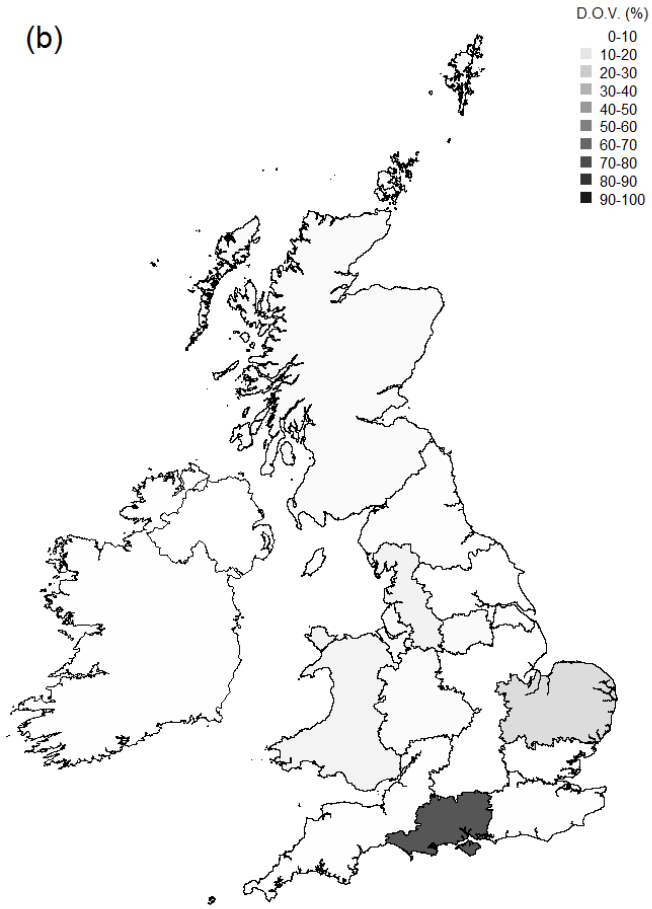




Diagnostic Origin Value at the Referral Cluster level

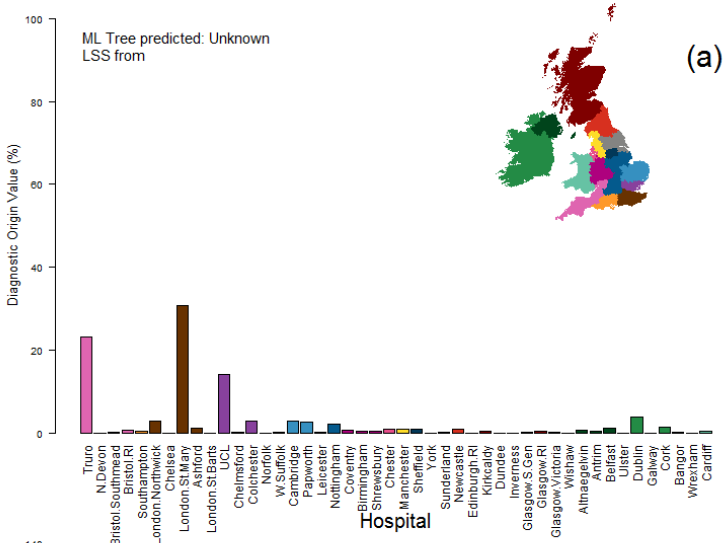
Isolate #936 (X7564_8.69) sampled from Southampton in 2010

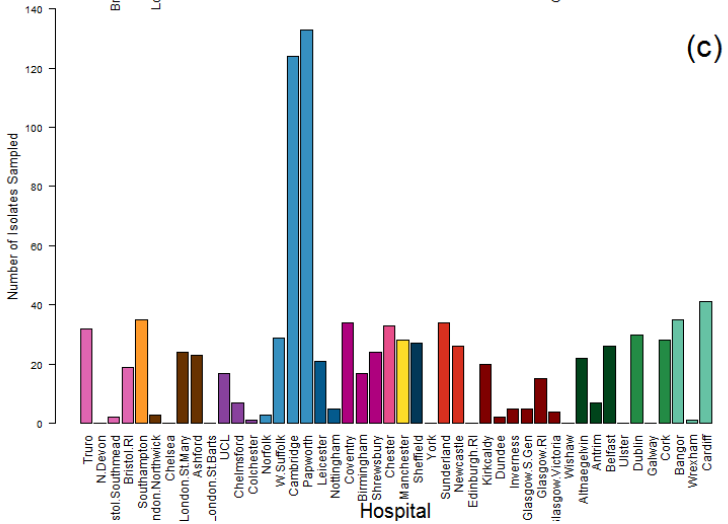


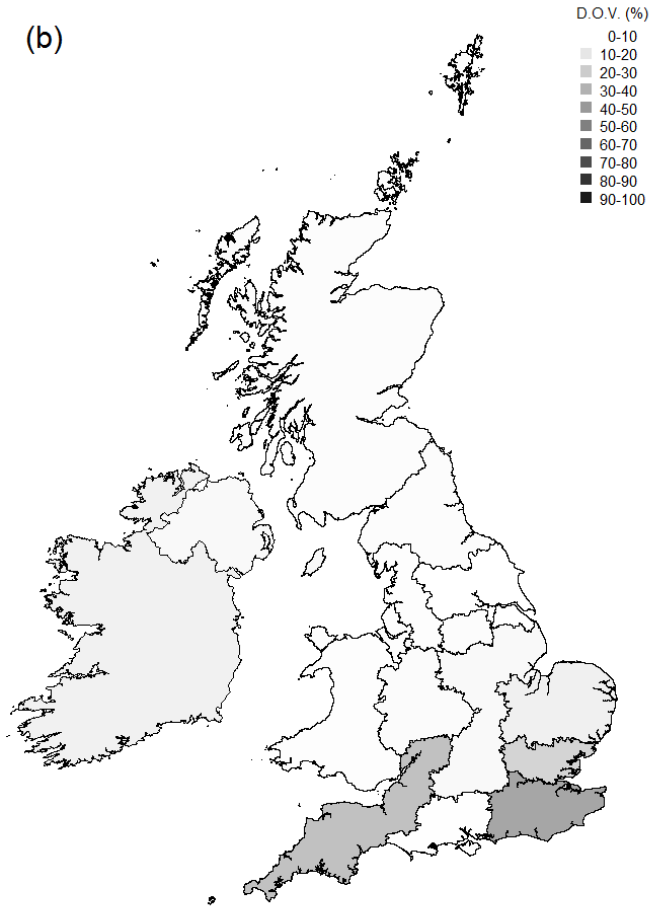


Diagnostic Origin Value at the Referral Cluster level

Isolate #943 (X7564_8.35) sampled from Chelsea in 2010

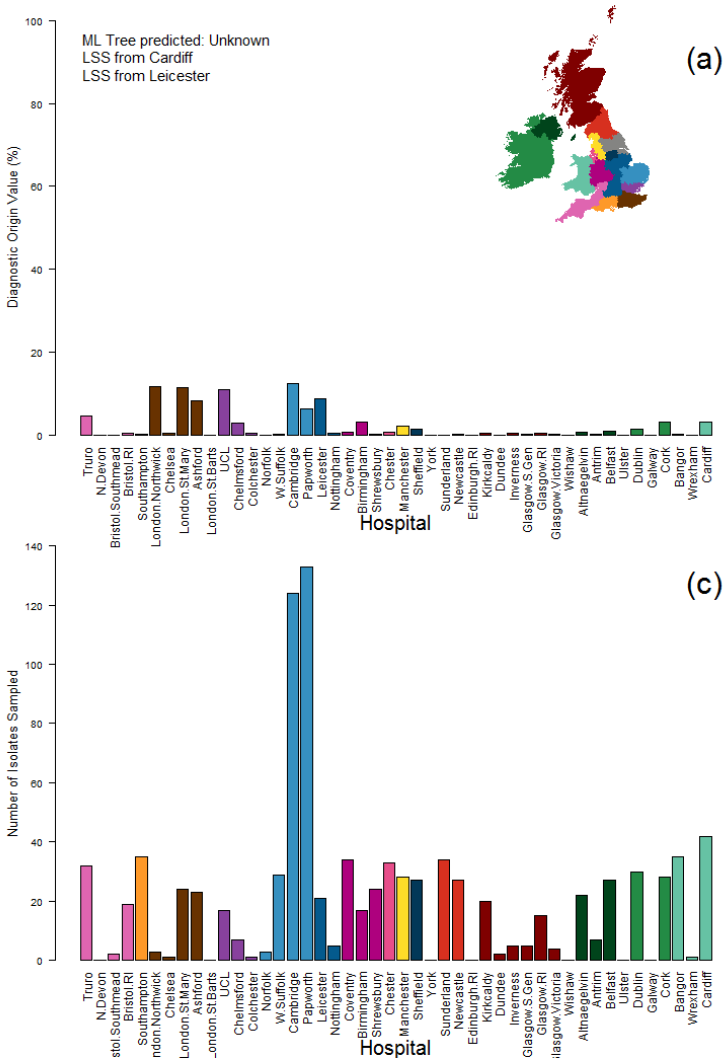


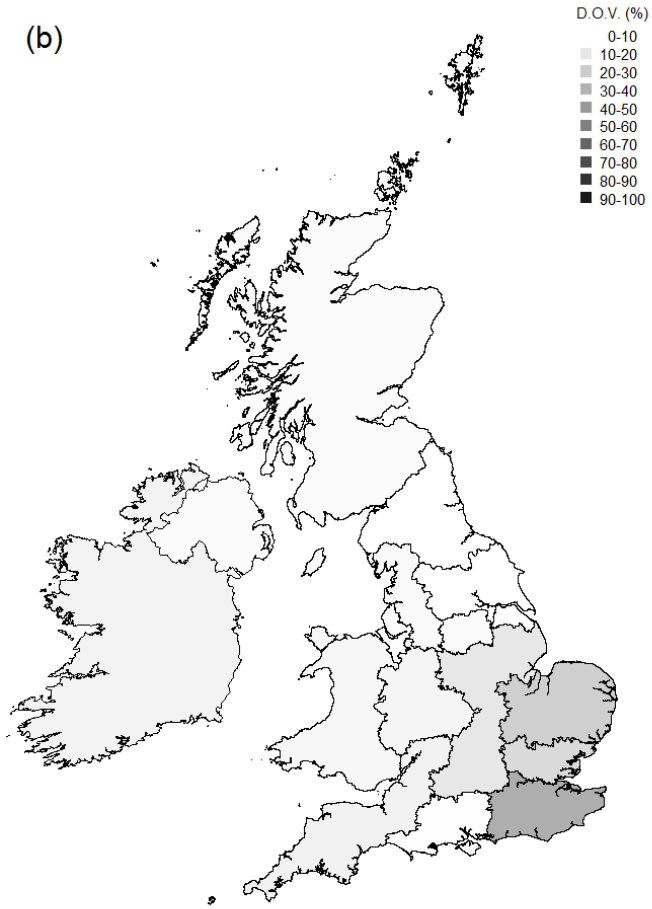




Diagnostic Origin Value at the Referral Cluster level

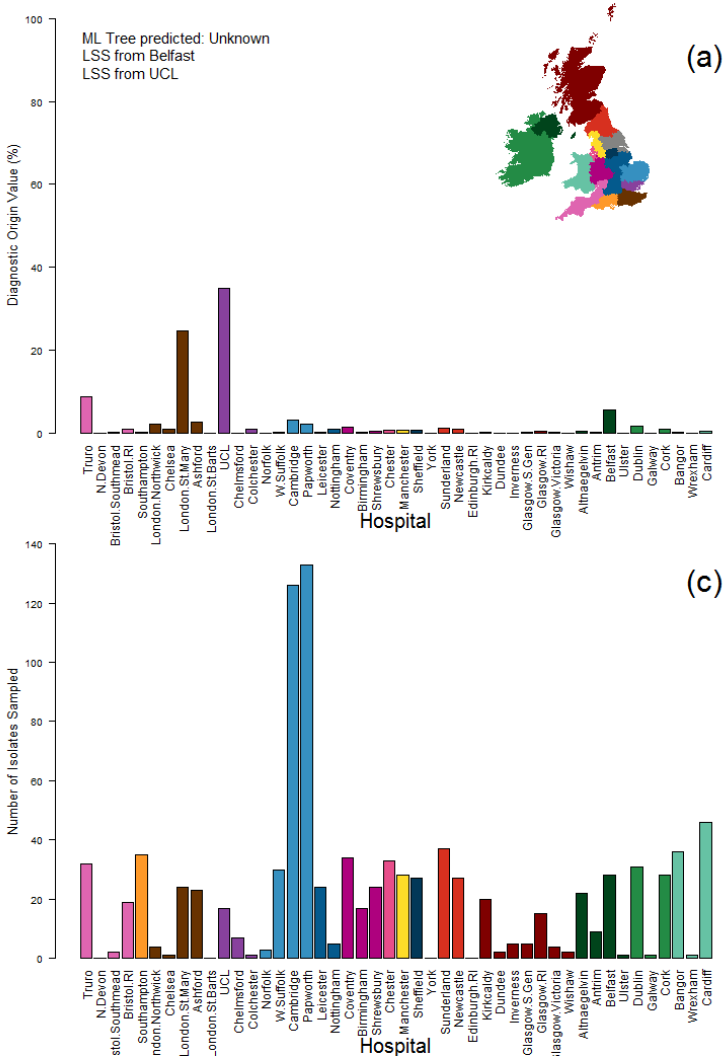
Isolate #947 (X7564_8.61) sampled from London.Northwick in 2010

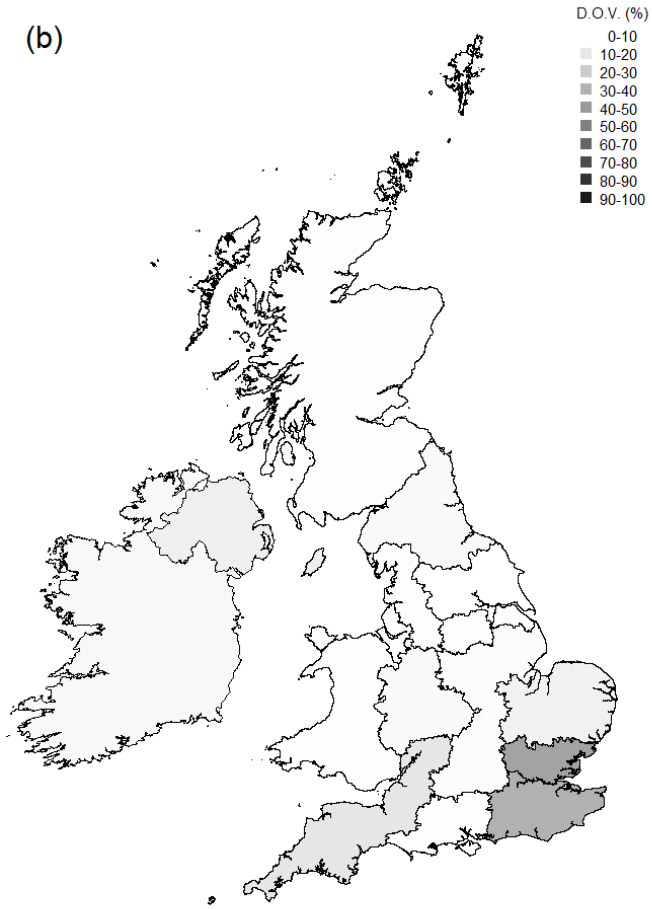




Diagnostic Origin Value at the Referral Cluster level

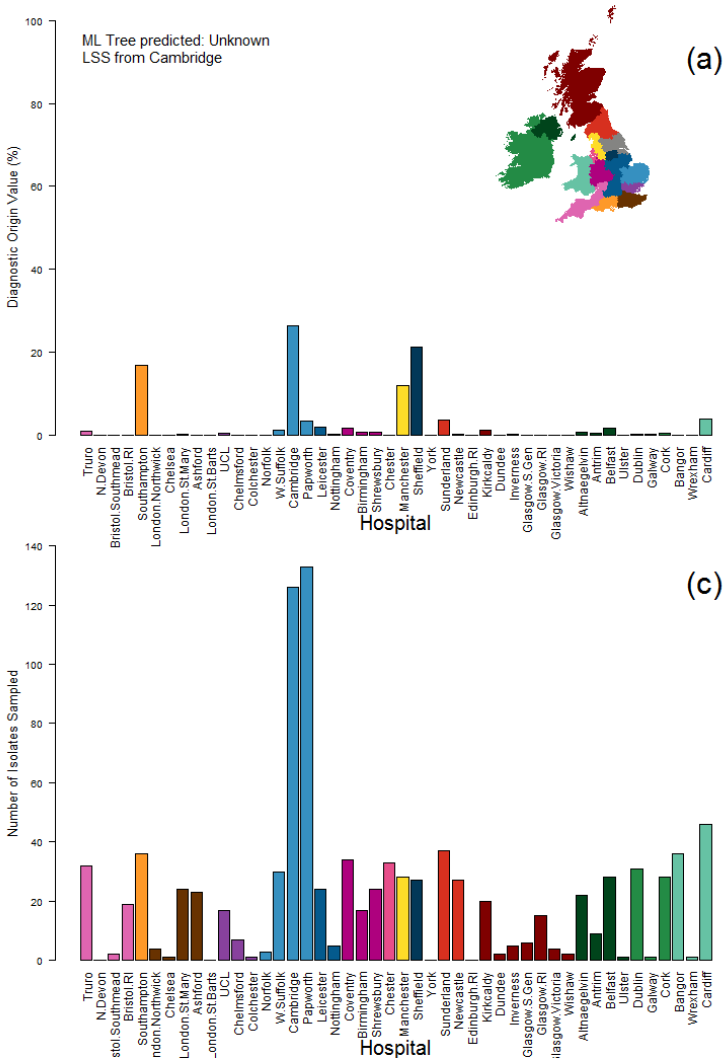
Isolate #970 (X7564_8.68) sampled from Glasgow.S.Gen in 2010

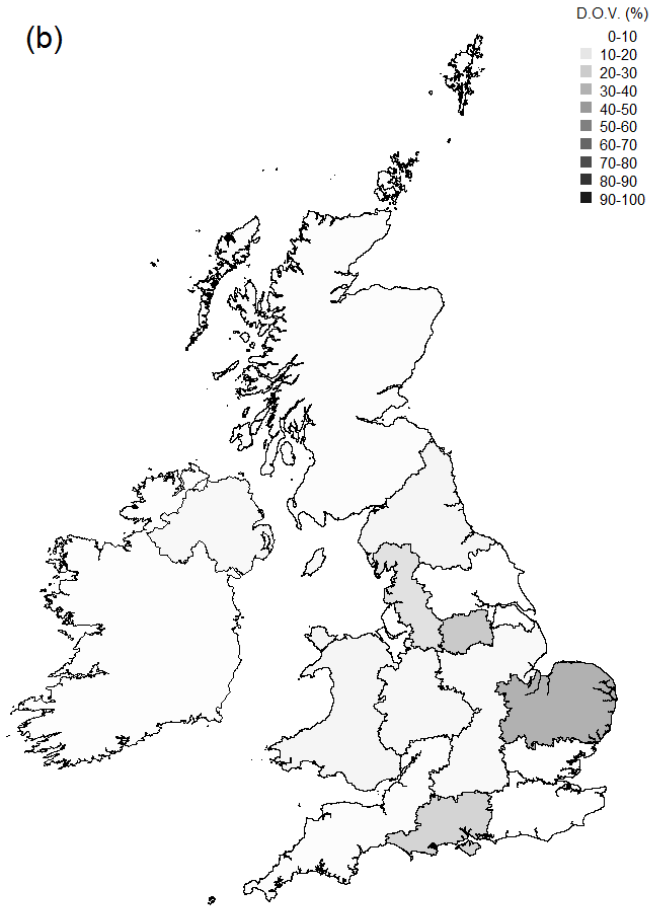




Diagnostic Origin Value at the Referral Cluster level

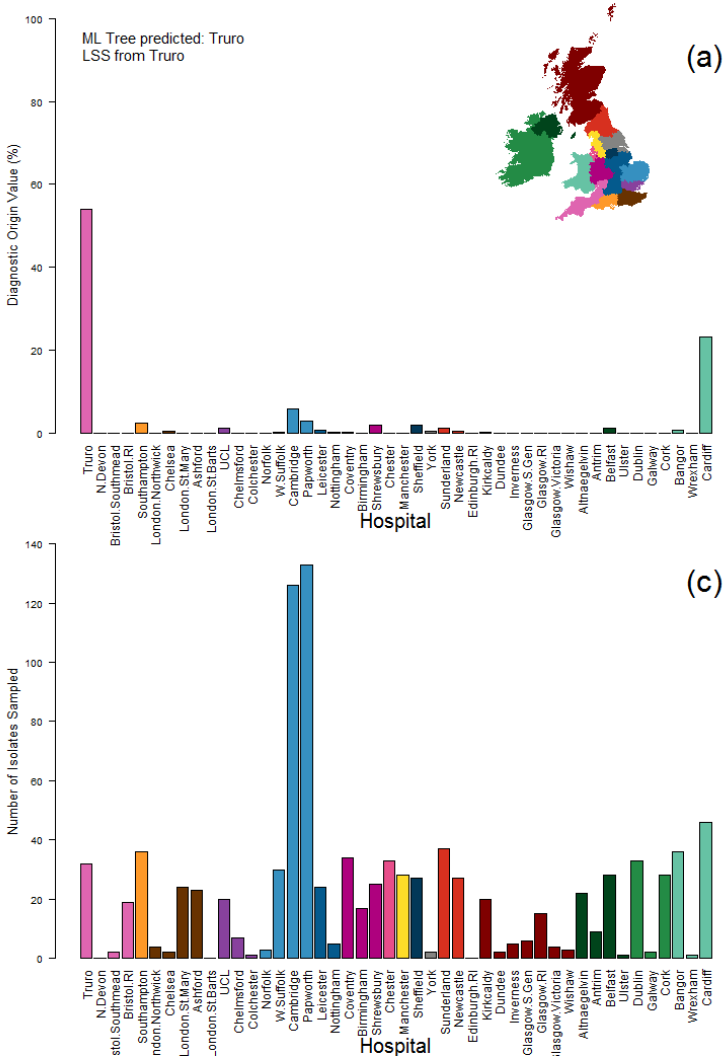
Isolate #972 (X7748_6.66) sampled from York in 2010

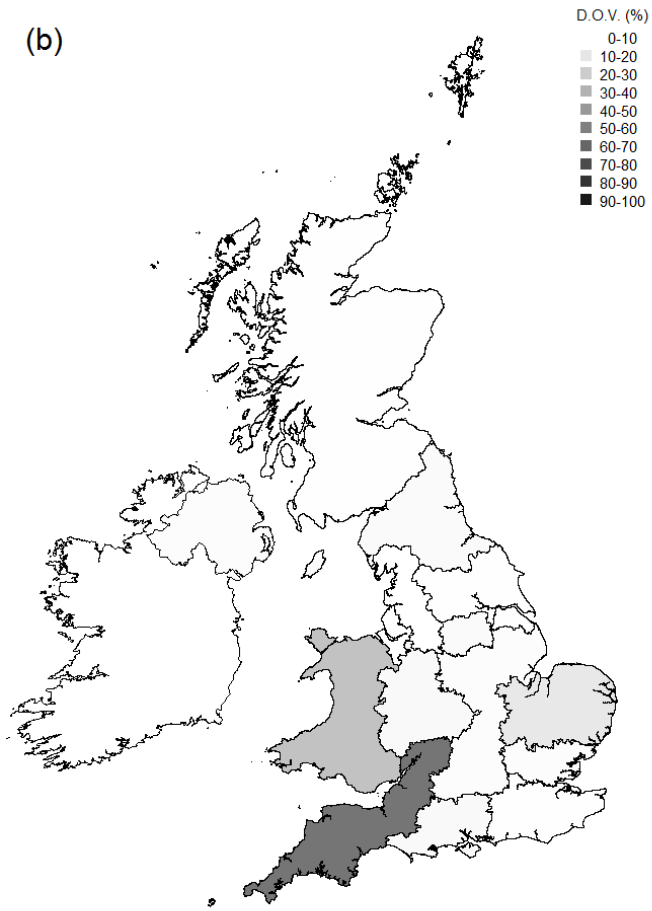




Diagnostic Origin Value at the Referral Cluster level

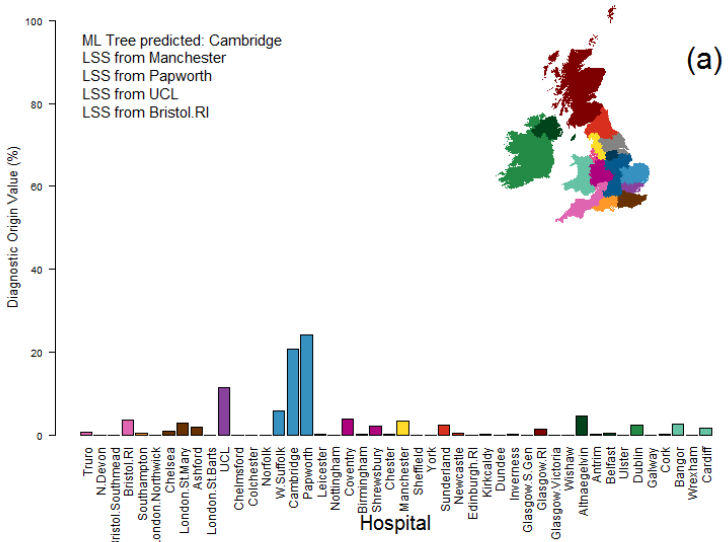
Isolate #983 (X7564_8.77) sampled from Truro in 2010

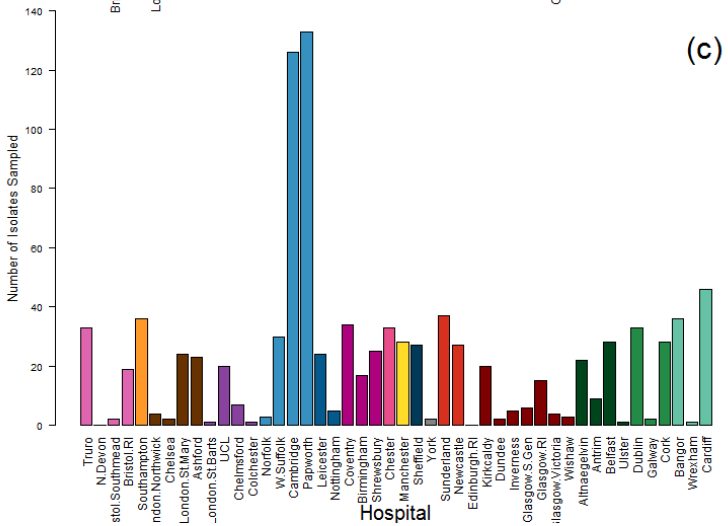


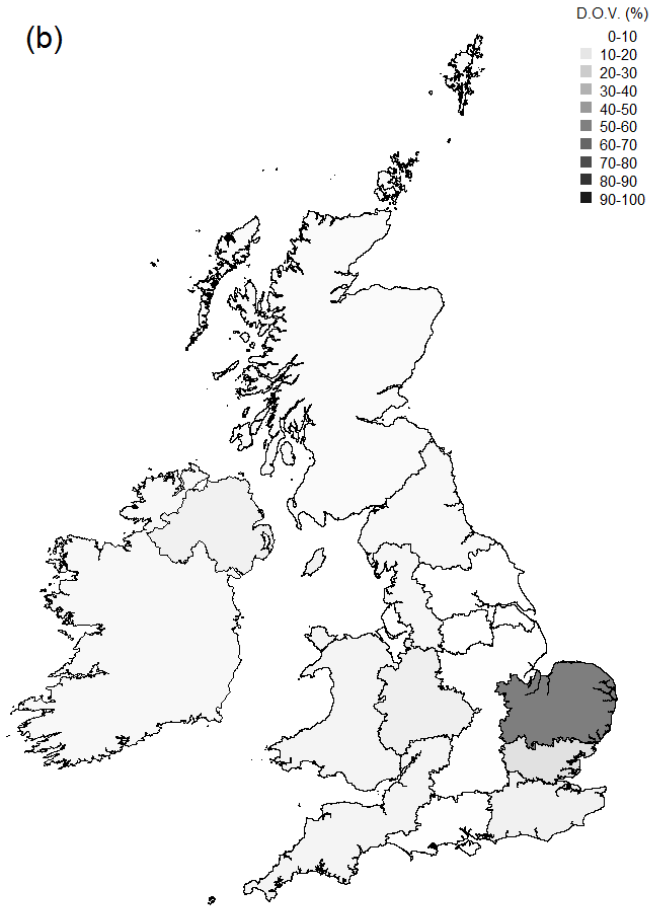


Diagnostic Origin Value at the Referral Cluster level

Isolate #985 (X7564_8.24) sampled from Cambridge in 2010

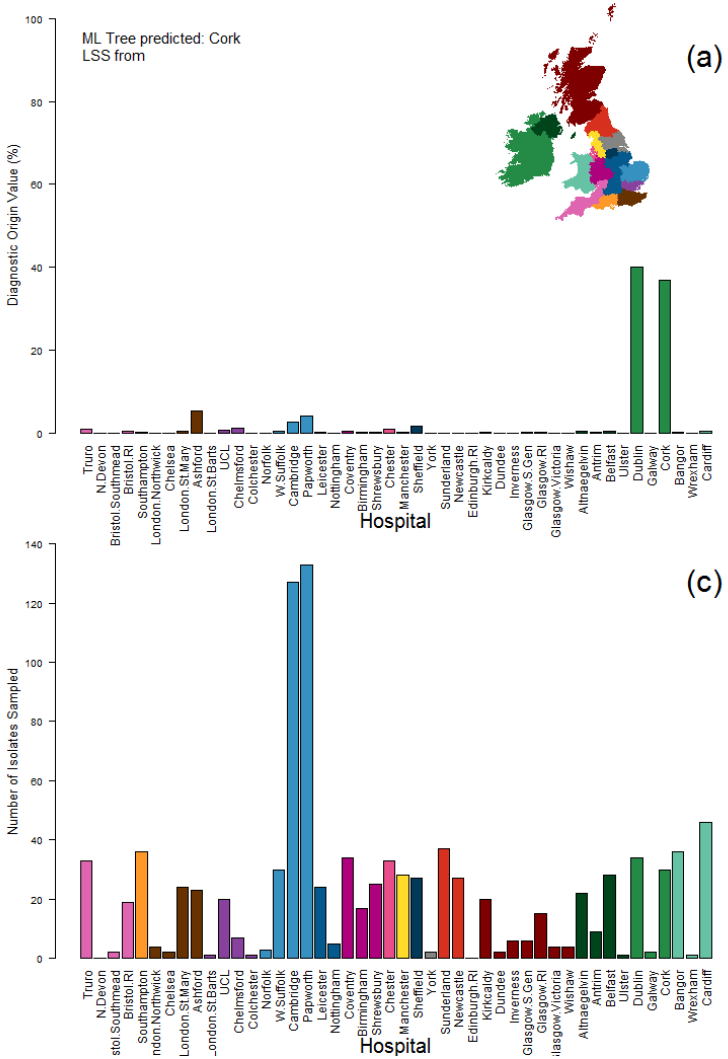


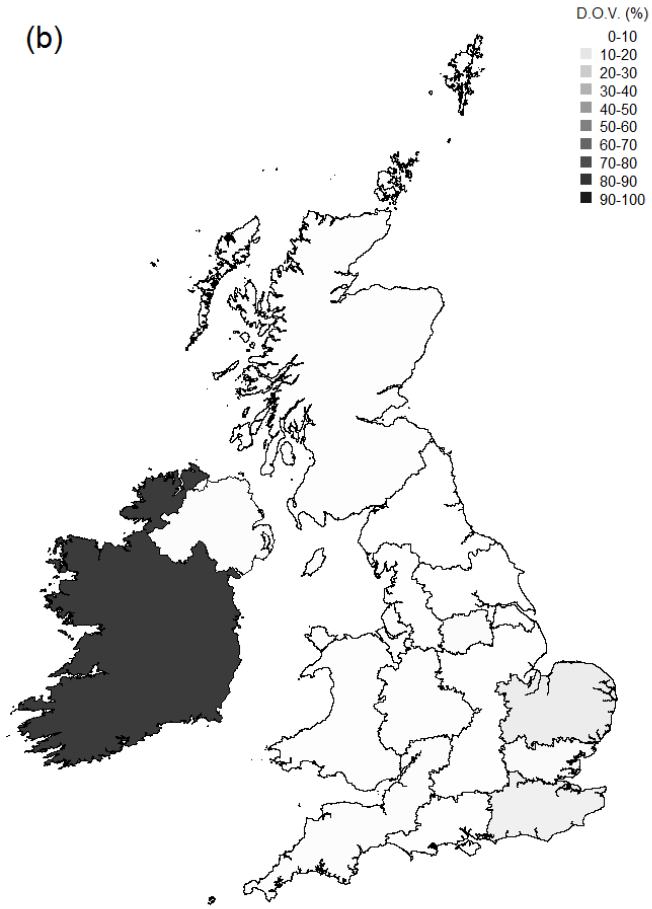




Diagnostic Origin Value at the Referral Cluster level

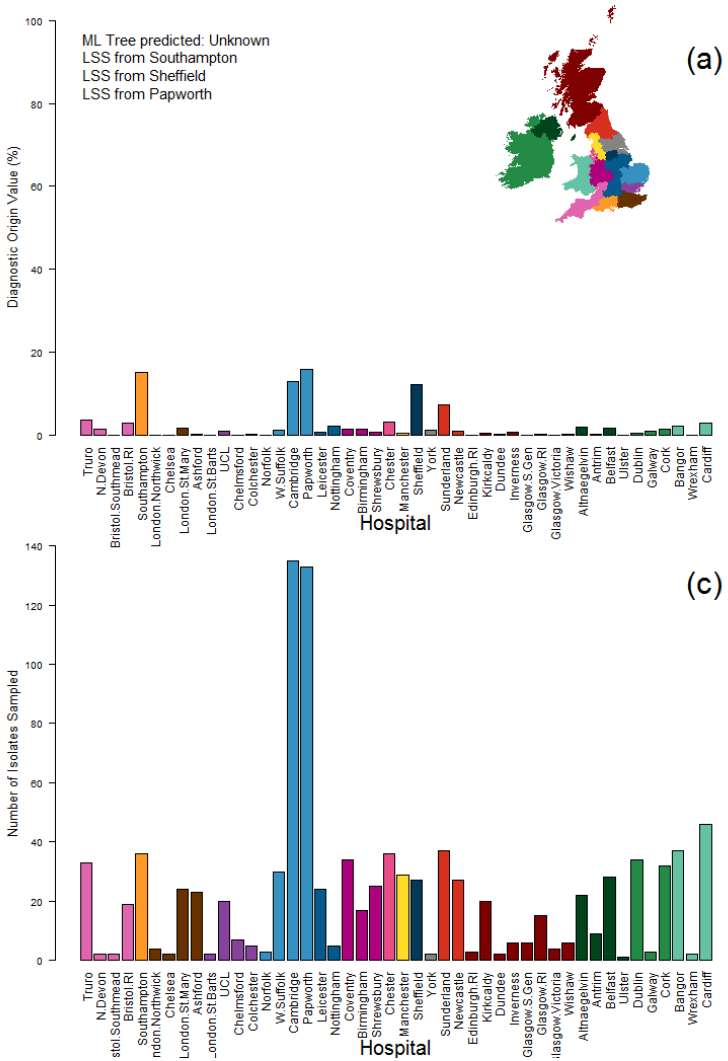
Isolate #991 (X7564_8.42) sampled from Cork in 2010

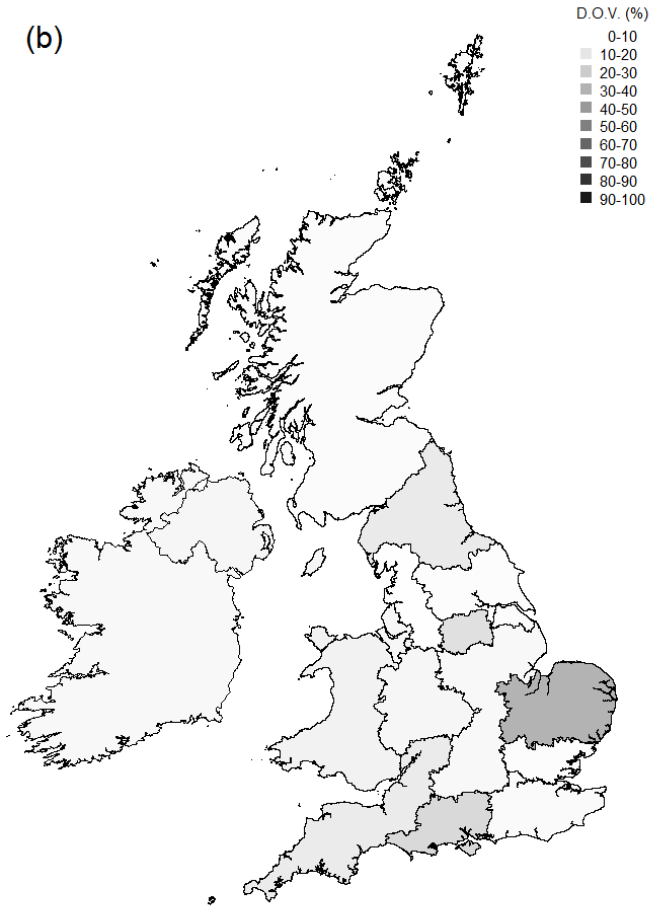




Diagnostic Origin Value at the Referral Cluster level

Isolate #1020 (X7915_6.14) sampled from Colchester in 2010





Diagnostic Origin Value at the Referral Cluster level

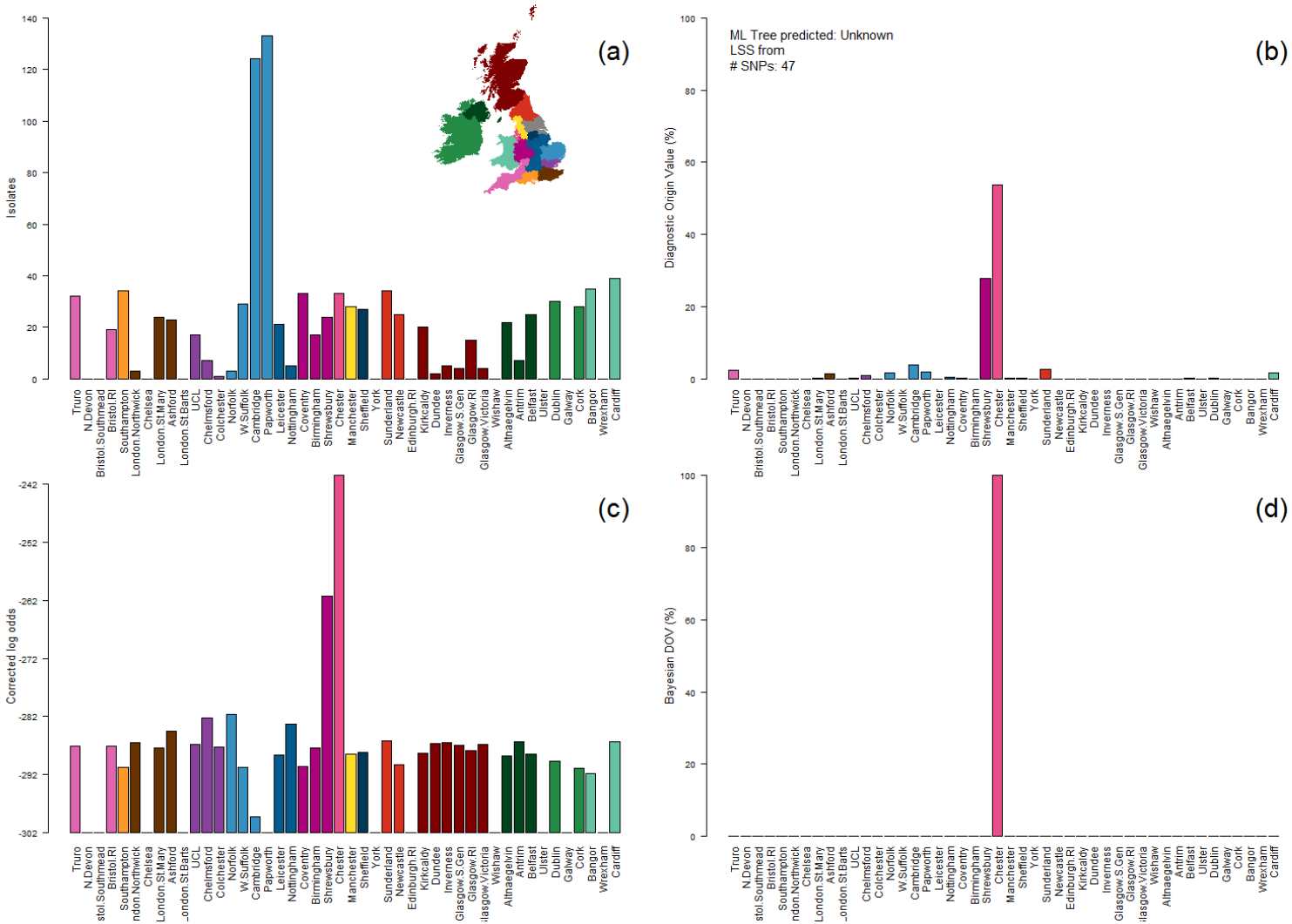
Appendix

F

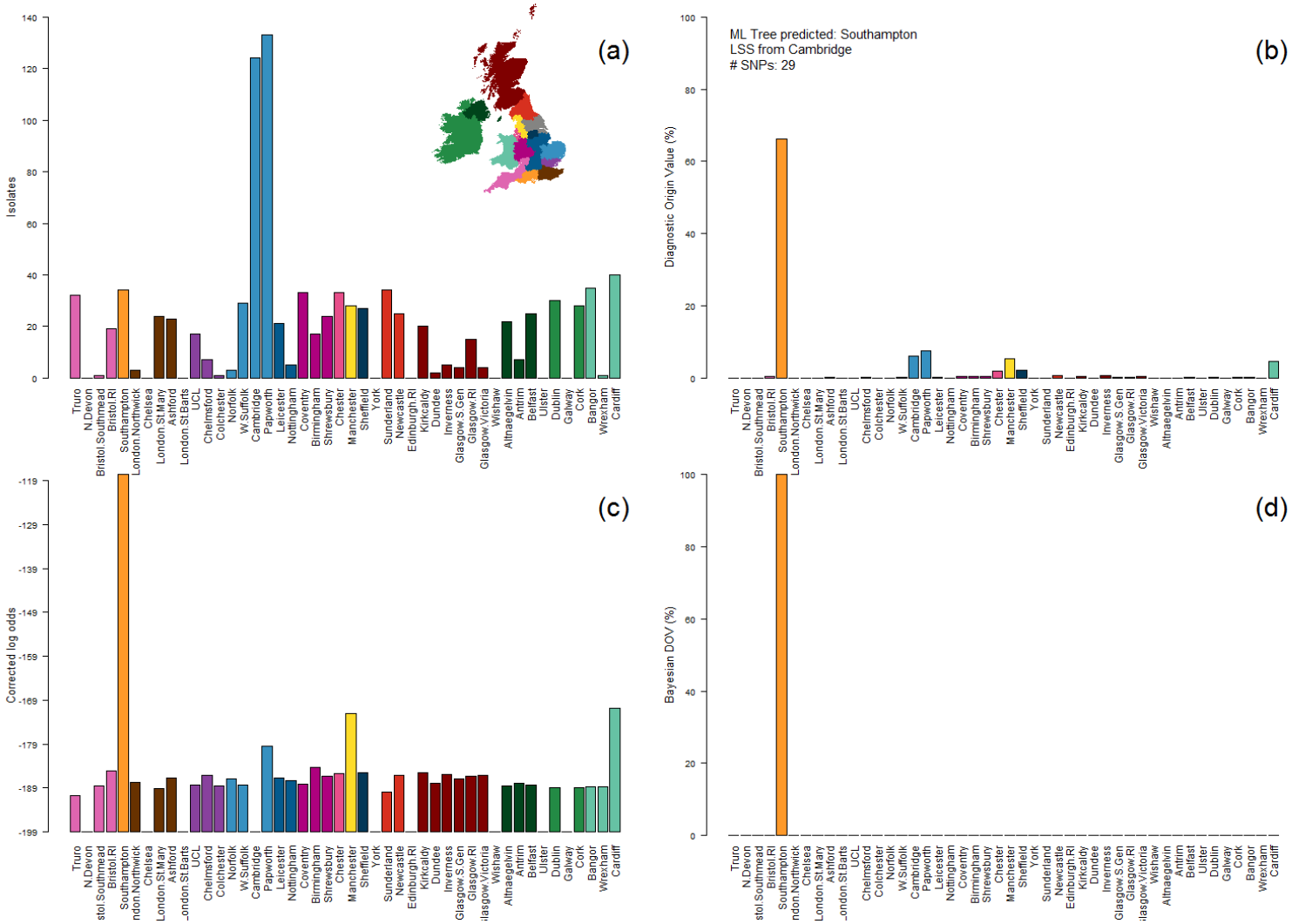
Here I provide a select number of examples from the 90 test isolates which were processed using the Bayesian classification approach, as described in Chapter 5. I chose these examples to exhibit the variety of output possible, and they are the same examples as those shown in Appendix E. Each of the following examples has the same layout, given here. The sampling effort of each hospital is shown in (a). Both the SnAPO (b) and the Bayesian method outputs (d) are included. I also include the log version of the Bayesian output (c). I also display any LSSs the isolate expresses, the number of SNPs the isolate harbours, and the origin location as predicted by TAPO in (b). In all plots the bars are coloured by Referral Cluster, provided as a graphical legend in (a). Brief descriptions of the following images are provided here:

- Isolate 933 shows the typical high peak in the Bayesian output, which concurs with one of the two high peaks in the SnAPO output.
- Isolate 936 shows the typical high peak in the Bayesian output which concurs with the SnAPO output. The log Bayesian output also appears uniform.
- Isolate 943 shows an ambiguous SnAPO output, though the Bayesian output concurs with the highest SnAPO peak.
- Isolate 947 shows a very unclear SnAPO output, and therefore the high peak seen in the Bayesian output could be suspect.
- Isolate 970 shows a discrepancy between the predicted SnAPO and Bayesian origin.
- Isolate 972 shows an ambiguous SnAPO output, with a mismatch between the SnAPO predicted origin and the Bayesian one.
- Isolate 983 shows concurrence between the two methods, however there are two possible origin locations with high values.
- Isolate 985 shows an ambiguous SnAPO output, though the Bayesian output concurs with the high peaks of the SnAPO output.
- Isolate 991 shows two possible SnAPO origins, with the Bayesian output concurring with the slightly lower of the two.
- Isolate 1020 shows an unclear SnAPO output, with an ambiguous Bayesian output.

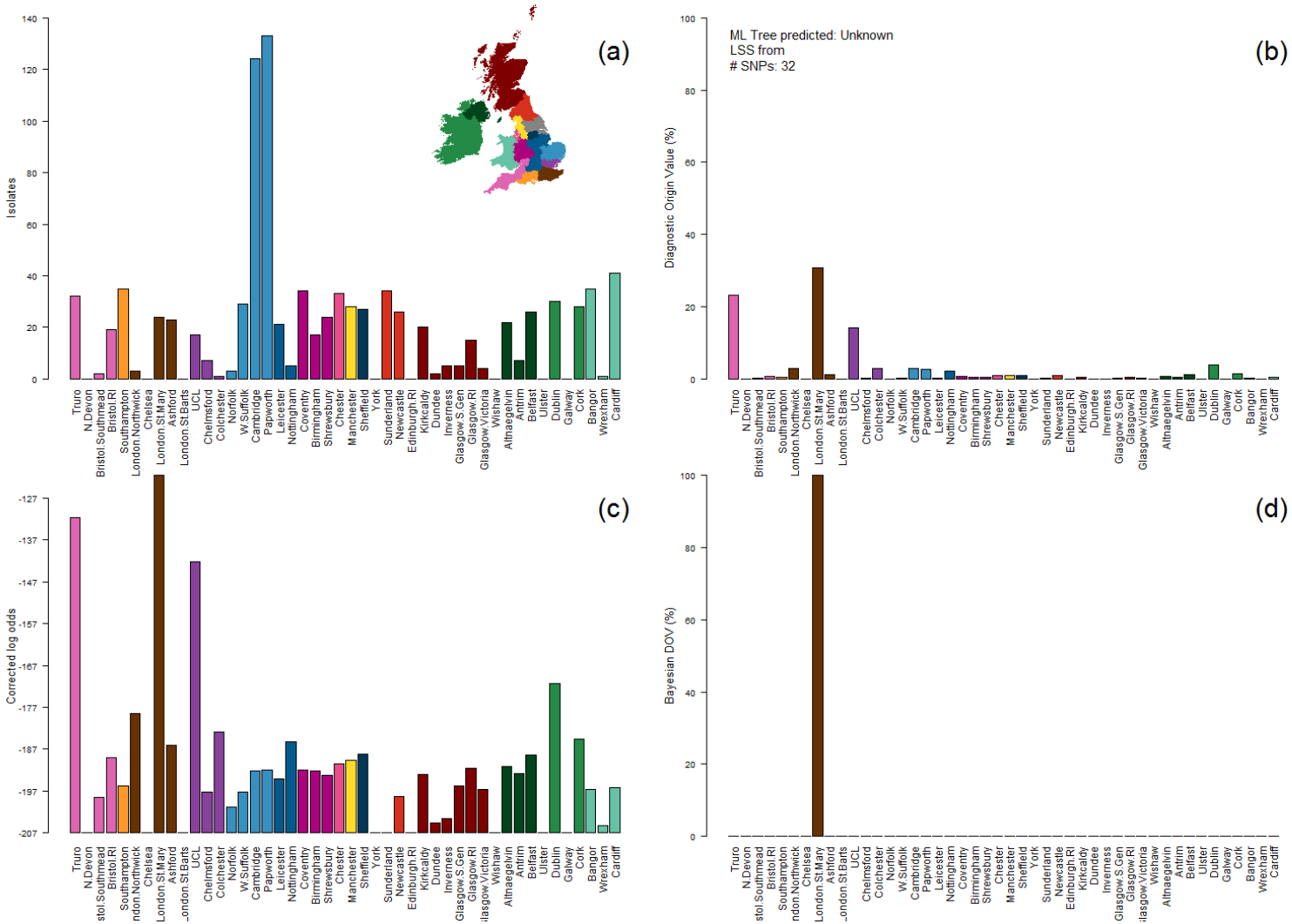
Isolate #933 (X7564_8.91) sampled from Wrexham in 2010



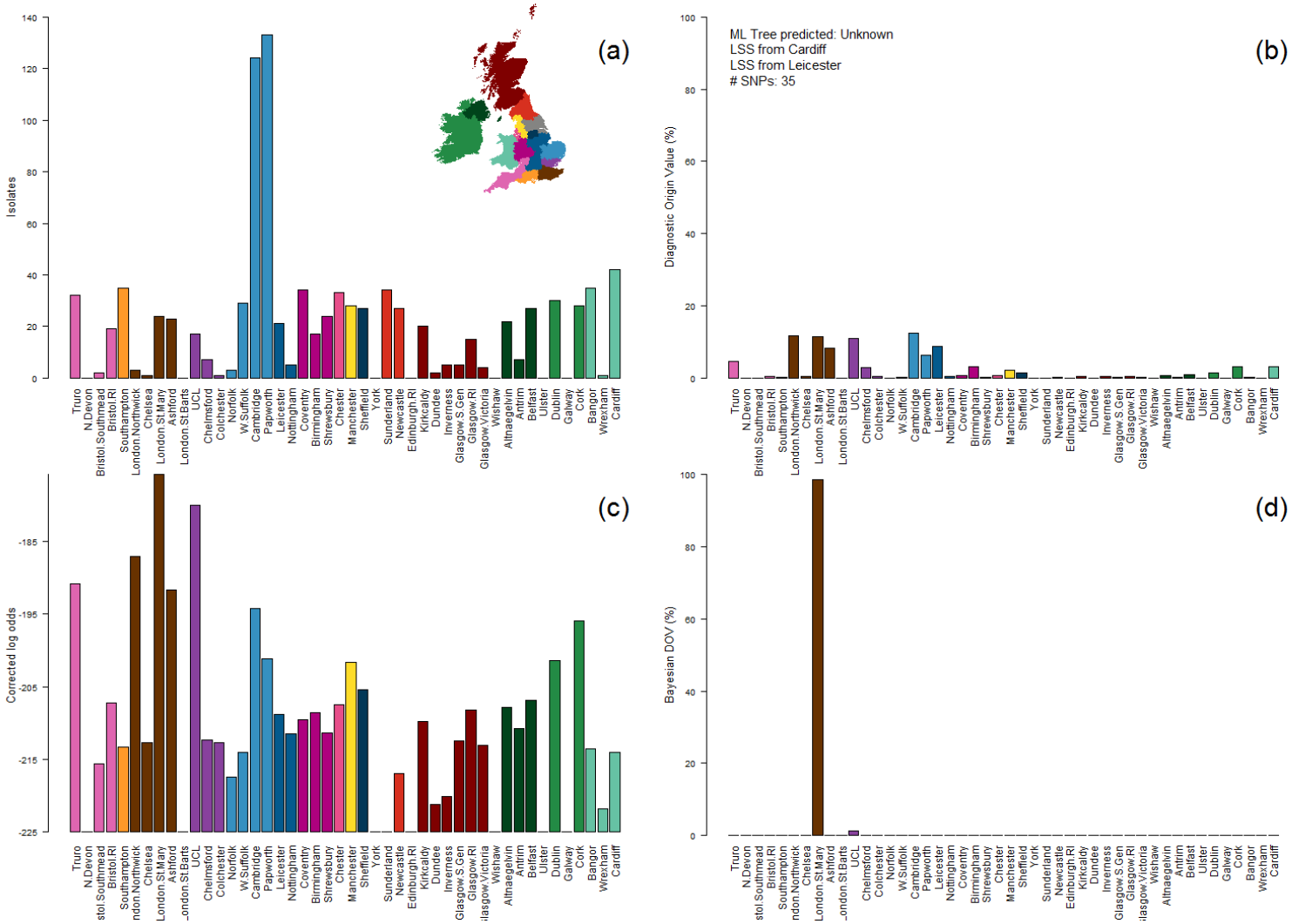
Isolate #936 (X7564_8.69) sampled from Southampton in 2010



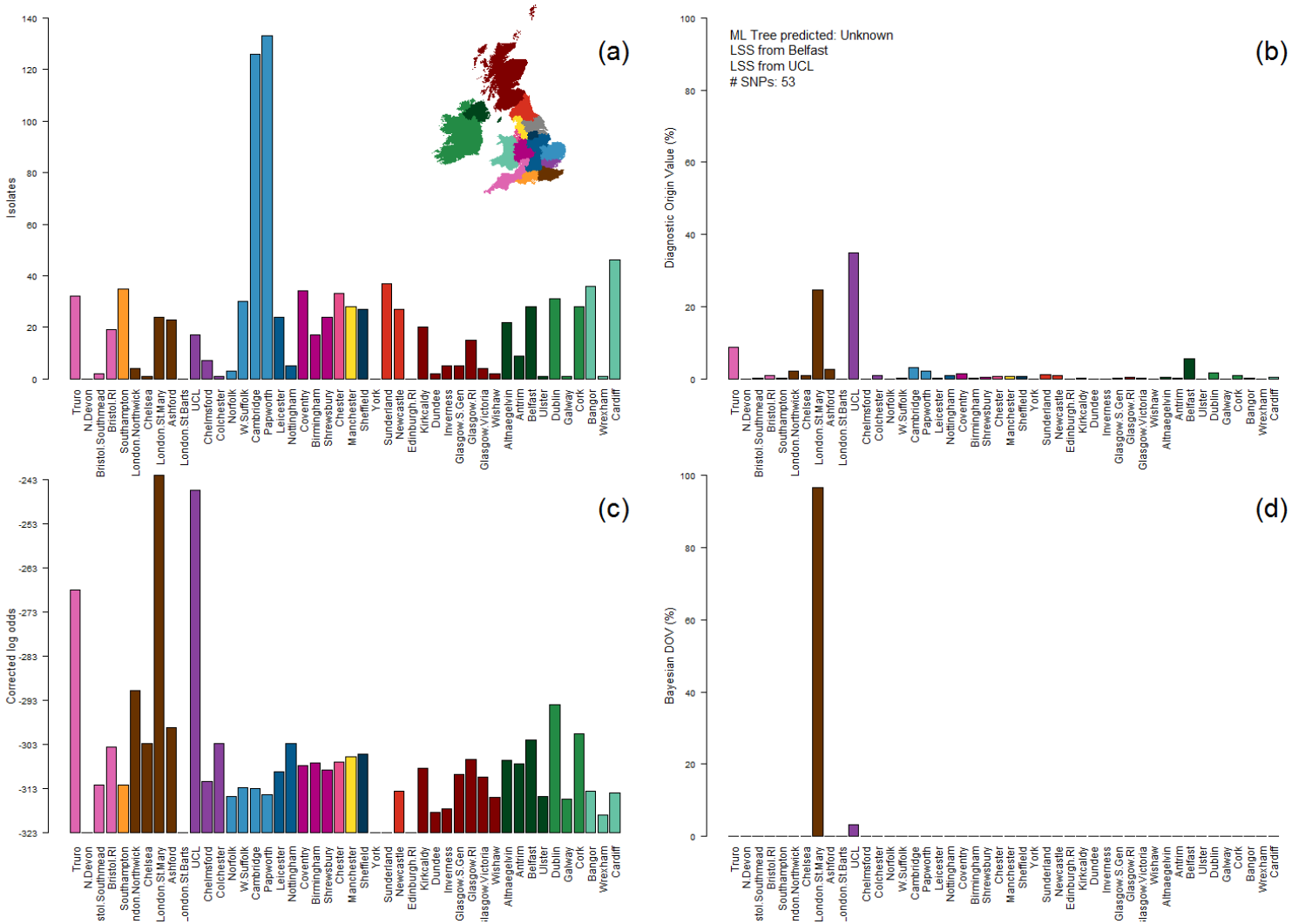
Isolate #943 (X7564_8.35) sampled from Chelsea in 2010



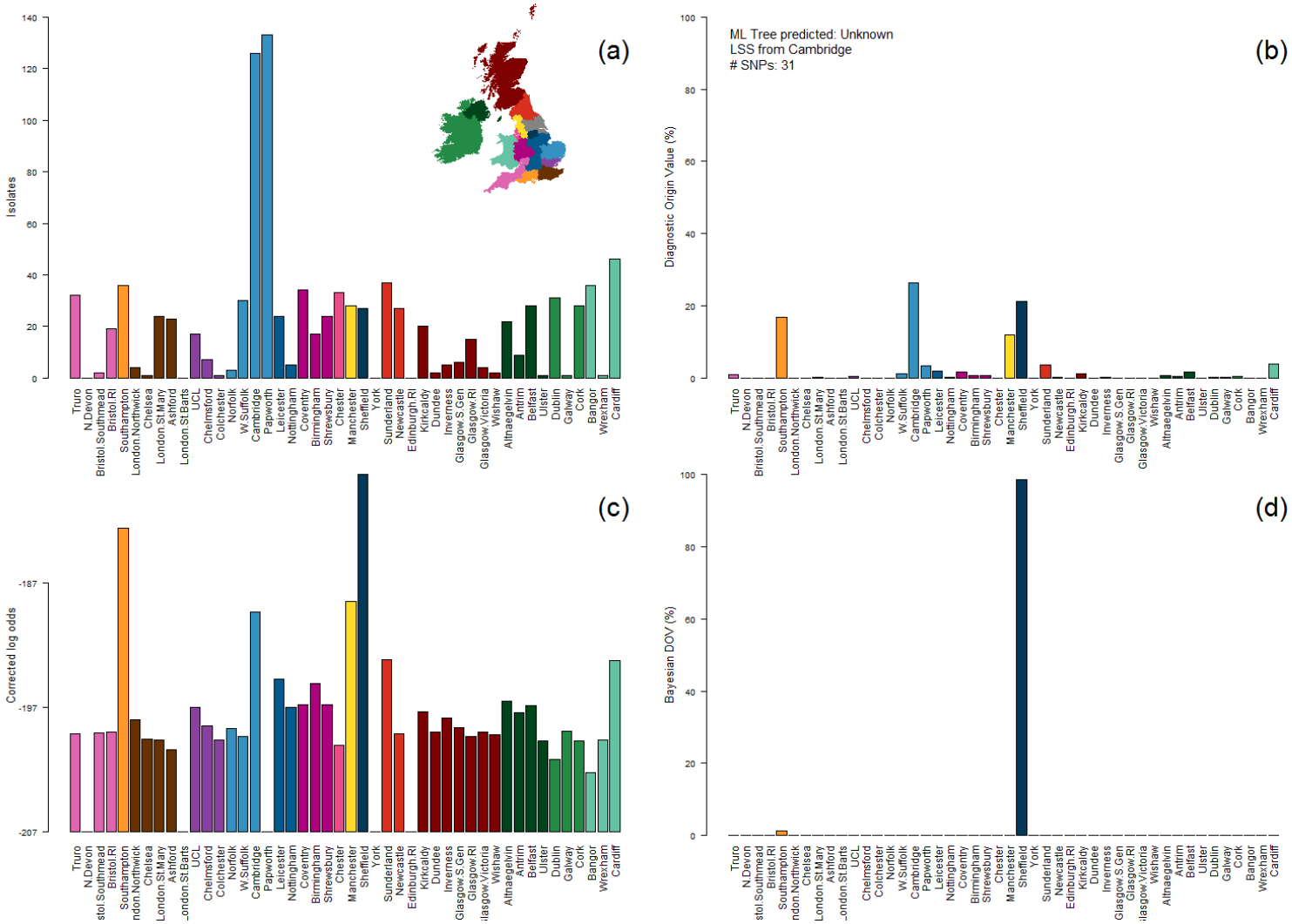
Isolate #947 (X7564_8.61) sampled from London.Northwick in 2010



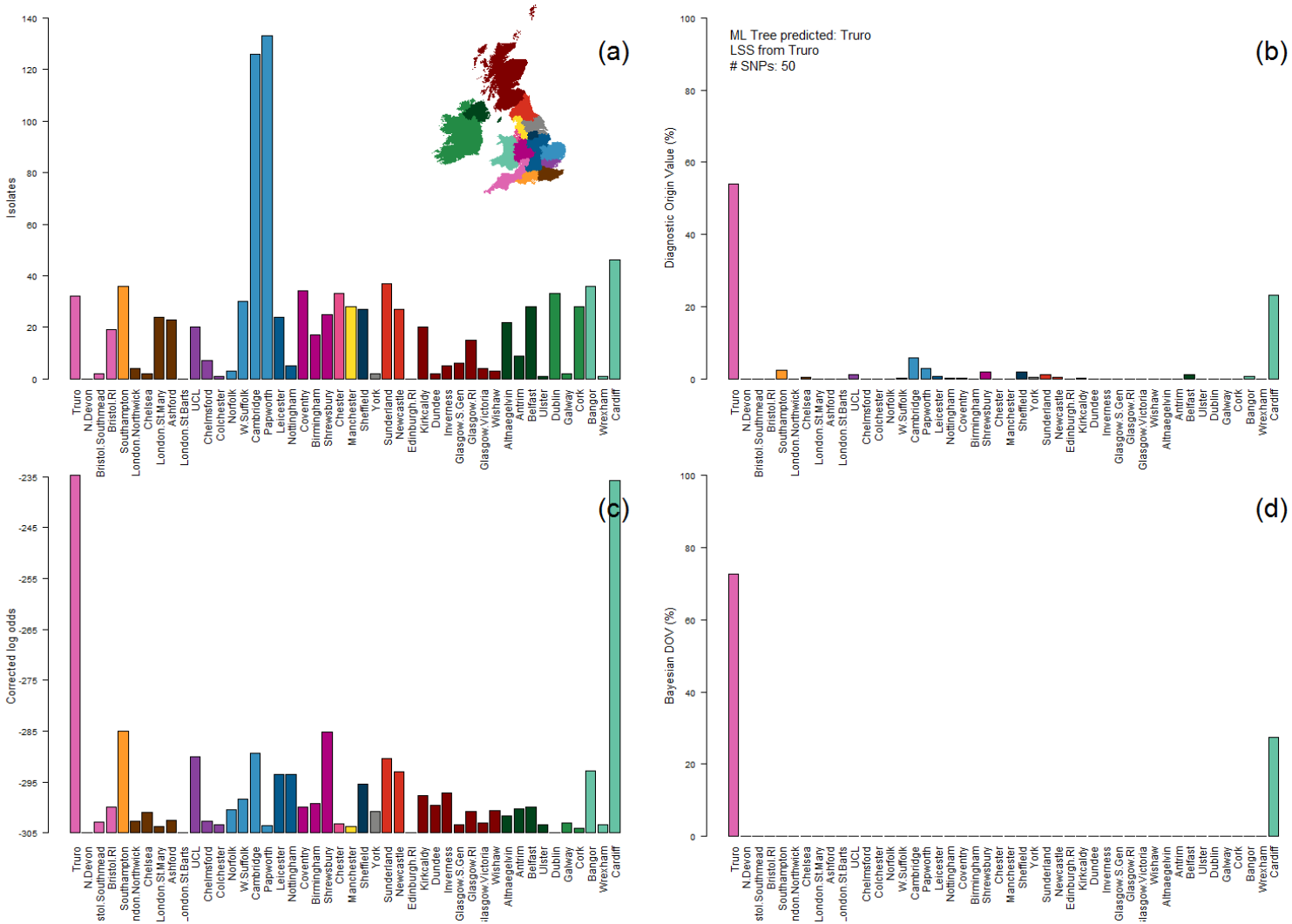
Isolate #970 (X7564_8.68) sampled from Glasgow.S.Gen in 2010



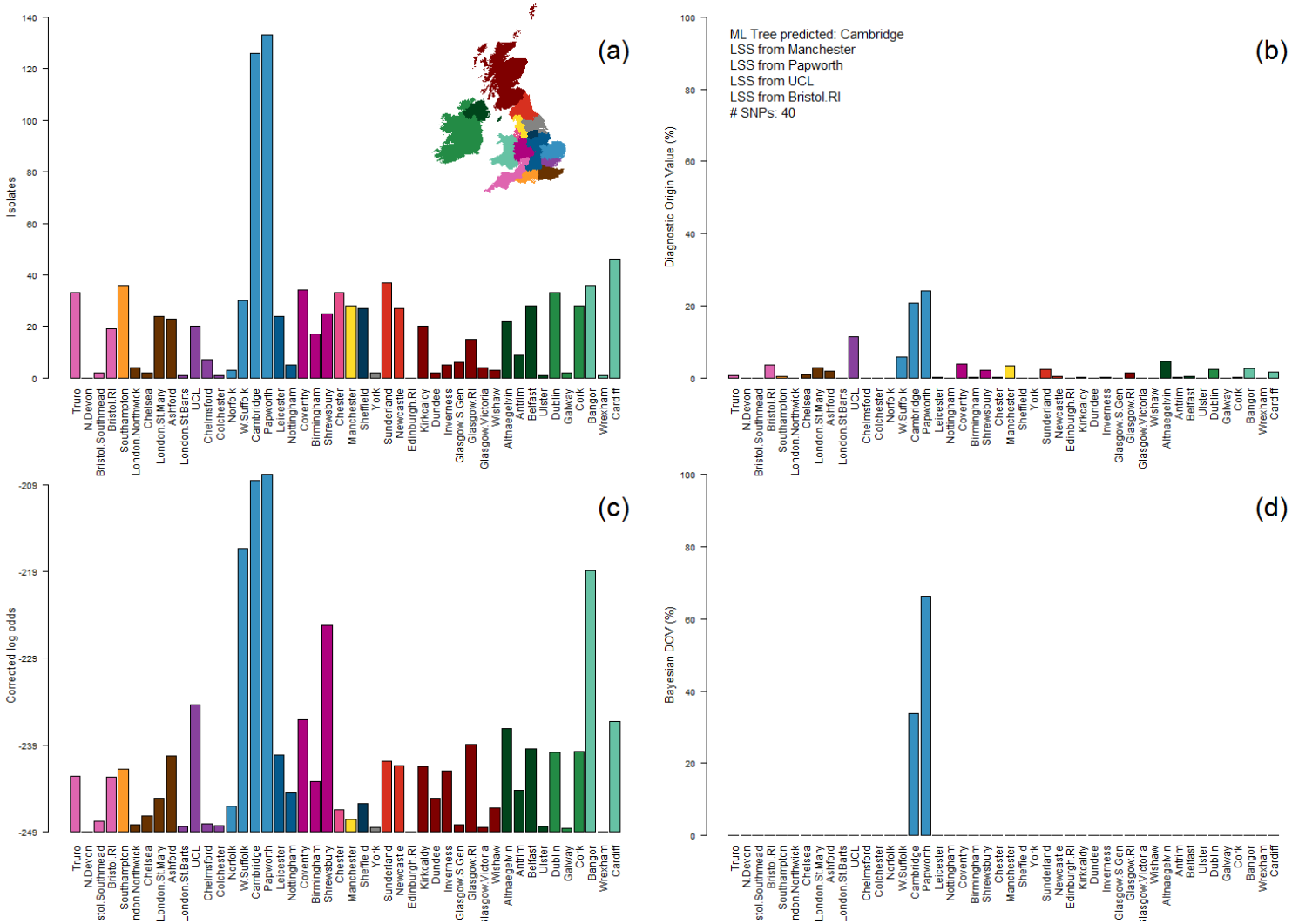
Isolate #972 (X7748_6.66) sampled from York in 2010



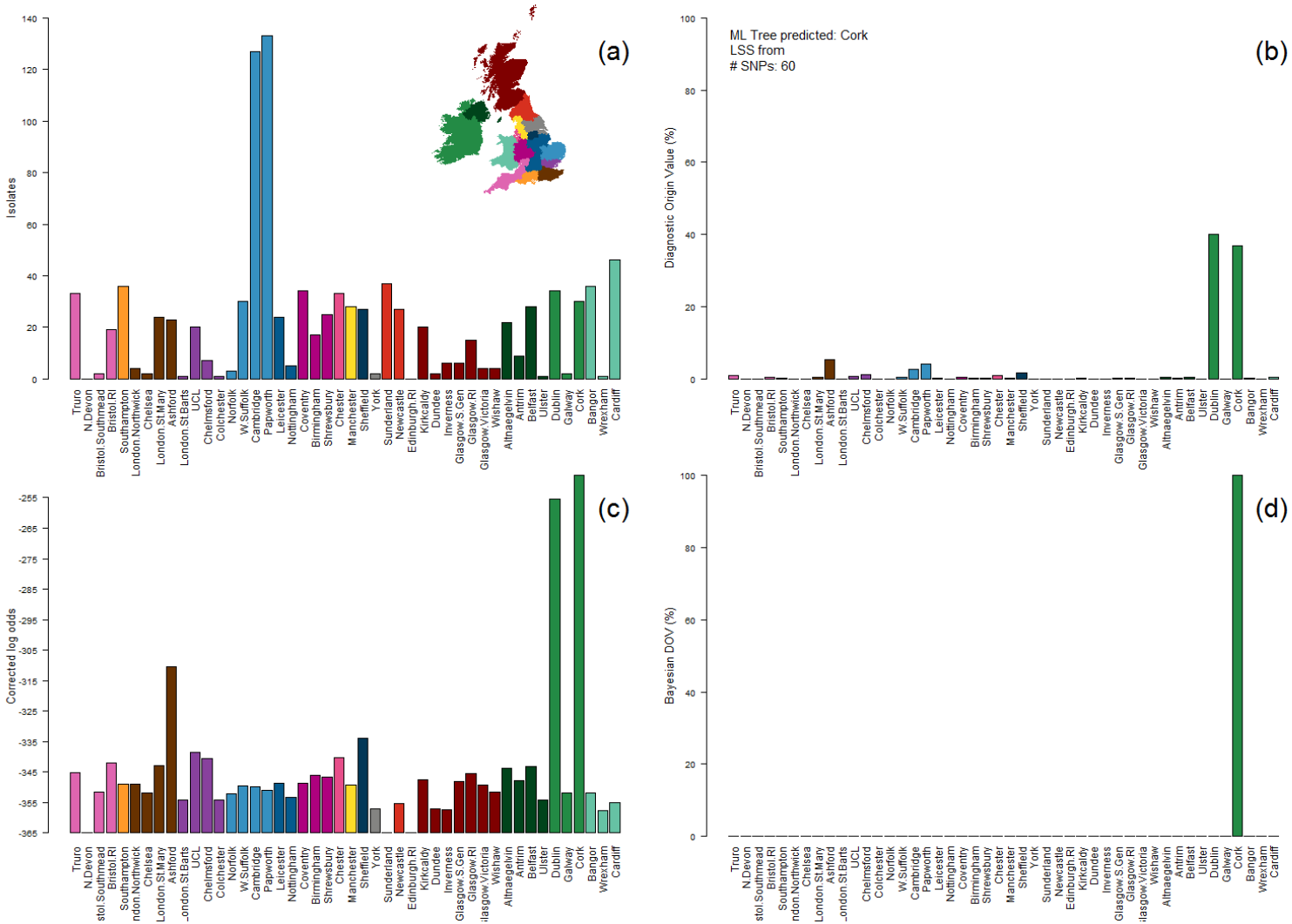
Isolate #983 (X7564_8.77) sampled from Truro in 2010



Isolate #985 (X7564_8.24) sampled from Cambridge in 2010



Isolate #991 (X7564_8.42) sampled from Cork in 2010



Isolate #1020 (X7915_6.14) sampled from Colchester in 2010